下一代数据库接口：LLM 基础的文本到 SQL 转换综述

紫金红 1，郑元 张庆刚 2，陈浩 2
董俊南 2. 黄飞然 1和小黄 2
1中国广州，暨南大学
香港理工大学，中国香港特别行政区hongzijin@stu2020.jnu.edu.cnyzheng.yuan@connect.polyu.hkqinggangg.zhang@connect.polyu.hksundaychenhao@gmail.comhanson.dong@connect.polyu.hkhuangfr@jnu.edu.cnxiaohuang@comp.polyu.edu.hk

从自然语言问题生成准确的SQL（文本到SQL）一直是一个长期的挑战，这主要是因为用户问题理解、数据库模式理解和SQL生成的复杂性。传统的文本到SQL系统，包括人工工程和深度神经网络，已经取得了实质性进展。随后，预训练语言模型（PLMs）被开发并用于文本到SQL任务，取得了令人鼓舞的性能。随着现代数据库变得更加复杂，相应的用户问题也变得更加具有挑战性，导致参数受限的PLMs产生错误的SQL。这就需要更复杂和定制化的优化方法，这反过来又限制了基于PLMs系统的应用场景。最近，大型语言模型（LLMs）在自然语言理解方面展现出显著的能力，随着模型规模的增加。因此，集成基于LLMs的实施可以为文本到SQL研究带来独特的机会、改进和解决方案。在这项调查中，我们提供了对基于LLMs的文本到SQL的全面回顾。具体来说，我们提出了一个技术挑战和文本到SQL发展过程的简要概述。然后，我们详细介绍了设计用于评估文本到SQL系统的数据集和指标。之后，我们系统地分析了基于LLMs的文本到SQL的最新进展。最后，我们讨论了这个领域仍然存在的挑战，并提出对未来研究方向的期望。

文本到SQL，大型语言模型，数据库，自然语言理解

I. INTRODUCTION

TEXT-TO-SQL是自然语言处理研究中的一项长期任务。它旨在将自然语言问题转换成数据库可执行的SQL查询。图1提供了一个基于大型语言模型（LLM基于）的文本到SQL系统的示例。给定一个用户问题，如

“你能告诉我历史上比赛次数最多的五个联赛的名称以及这些联赛中分别进行了多少场比赛吗？”LLM接受这个问题及其对应的数据库模式作为输入，然后生成一个

SQL查询作为输出。这个SQL查询可以在数据库中执行，以检索相关内容来回答用户的问题。上述系统构建了一个使用LLMs的数据库自然语言接口（NLIDB）。由于SQL仍然是最广泛使用的编程语言之一，拥有超过

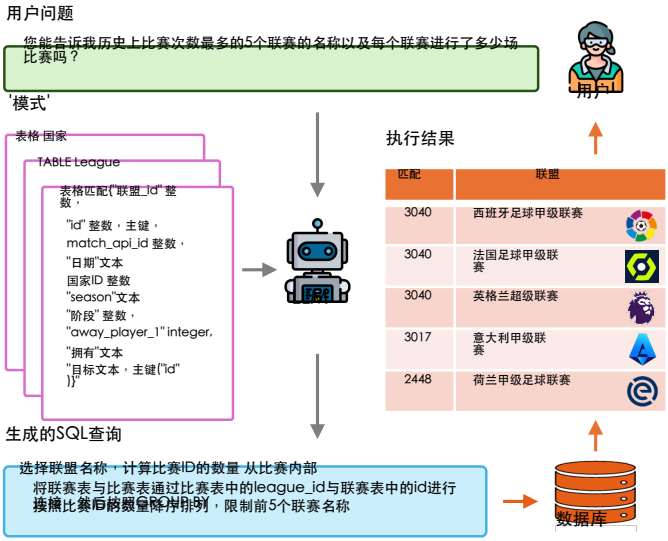

图1：从BIRD数据集中选取的基于LLM的文本到SQL的例子。一个用户提出了一个关于足球联赛的问题。LLM将这个问题和其对应数据库的模式作为输入，然后生成一个

SQL查询作为输出。该SQL查询可以在数据库中执行，并检索内容“比赛次数最多的5个联赛”，以回答用户的问题。

在他们的工作中，有一半（51.52%）的专业开发人员使用SQL，值得注意的是，只有大约三分之一（35.29%）的开发人员接受过系统培训。NLIDB使非技术用户能够像专业数据库工程师一样访问结构化数据库[1, 2]，并且还加速了人机交互[3]。此外，在研究热点中

大型语言模型（LLMs）通过将数据库中的真实内容整合进来，可以提供一种潜在的解决方案，以解决普遍存在的幻觉问题，填补知识空白。

LLMs [6]。文本到

SQL 引发了一系列关于其与 LLMs 集成和优化的研究 [7-10]；因此，基于 LLM 的文本到 SQL 转换仍然是 NLP 和数据库社区内高度讨论的研究领域。

通讯作者。

您提供的链接指向的是一个SurveyMonkey问卷调查，但该网页为空白网页，没有提供任何内容。因此，无法为您提供翻译服务。根据URL推测，该链接可能是一个针对2023年度的调查问卷，由Stack Overflow发起。Stack Overflow是一个流行的程序员问答网站，经常进行此类调查以了解开发者社群的技术趋势、偏好和工作习惯等。就专业术语而言，此类调查可能会涉及多种技术和行业专有名词，如“编程语言”（programming language）、“框架”（framework）、“数据库”（database）等。此外，还可能包含与开发者工作环境相关的术语，如“远程工作”（remote work）、“开源软件”（open-source software）

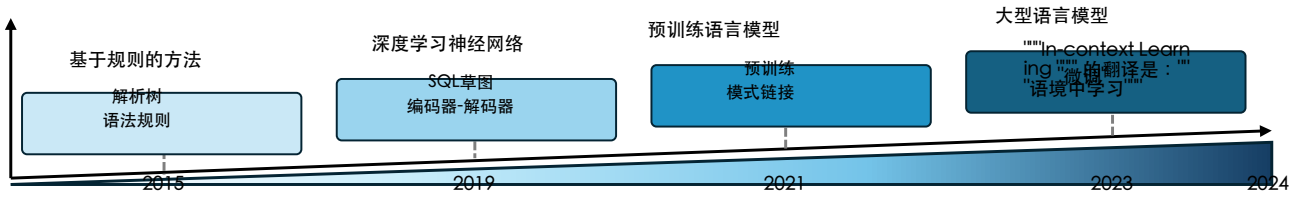

图2：从实现范式的视角，文本到SQL研究的演变过程草图。每个阶段都展示了两种代表性的实现技术。各阶段的时间节点并不完全准确；我们根据每种实现范式的代表性作品的发布时间来设定每个时间戳，前后误差大约一年。格式灵感来源于[29]。

先前的研究在实现文本到SQL转换方面取得了显著进展，并经历了漫长的演化过程。早期的努力主要基于精心设计的规则和模板[11]，特别适合简单的数据库场景。近年来，由于规则基础方法带来的沉重劳动力成本[12]以及数据库环境日益增加的复杂性[13-15]，为每种场景设计一个规则或模板变得越来越困难且不切实际。深度神经网络的发展推进了文本到SQL的进步[16, 17]，它可以自动学习从用户问题到其相应SQL的映射[18, 19]。随后，具有强大语义解析能力预训练语言模型（PLMs）已成为文本到SQL系统的新范式[20]，将它们的表现提升到新的水平[21-23]。对基于PLM的优化进行增量研究，如表内容编码[19, 24, 25]和预训练[20, 26]，进一步推进了这一领域。最近，通过上下文学习(ICL)[8]和微调(FT)[10]范式实现文本到SQL的LLM基础方法，以精心设计的框架和比PLMs更强的理解能力达到了最先进的准确度。

基于大模型的文本生成的整体实现细节

SQL可以分为3个方面：1. 问题理解

翻译结果：1. NL问题：NL问题是用户意图的语义表示，相应的生成SQL查询应与其对齐；2.模式理解：模式提供了数据库的表和列结构，文本到SQL系统需要识别与用户问题匹配的目标组件；3. SQL生成：这涉及整合上述解析，然后预测正确的语法以生成可执行的SQL查询，可以检索所需答案。LLMs已被证明能够实现良好的基础实施[7,27]，受益于更强大的语义解析能力，这是由更丰富的训练语料库启用的[28,29]。关于增强LLMs进行问题理解[8,9]、模式理解[30,31]和SQL生成[32]的进一步研究正在不断增加。

尽管在文本到SQL研究中取得了显著进展，但仍存在一些挑战阻碍着健壮且通用的文本到SQL系统的发展。近年来的相关研究已经对深度学习方法中的文本到SQL系统进行了调查，并提供了对以往研究的洞察。

深度神经网络和基于PLM的研究[2, 29, 74]。在这项调查中，我们旨在跟进最新的进展，并提供一个全面的综述，介绍当前最先进的模型和方法在基于LLM的文本到SQL转换中的应用。我们从介绍与文本到SQL转换相关的基本概念和挑战开始，强调这项任务在各个领域的重要性。然后，我们深入探讨了文本到SQL系统实现范式的演变，讨论了这个领域的关键进展和突破。概述之后，我们详细介绍并分析了最近在集成LLMs的文本到SQL方面的进展。具体来说，我们的调查内容涵盖了一系列与

基于大型语言模型的文本到SQL，包括：

数据集和基准测试：我们详细介绍了用于评估基于LLM的文本到SQL系统的常用数据集和基准测试。我们讨论了它们的特征、复杂性以及它们为文本到SQL开发和评估所带来的挑战。

• 评估指标：我们展示了用于评估基于大型语言模型（LLM）的文本到...性能的评估指标。

SQL系统包括基于内容匹配和基于执行的范式。接下来，我们简要介绍每种度量标准的特征。

• 方法和模型：我们系统分析了基于大型语言模型（LLM）的文本到SQL的不同方法和模型，包括上下文学习以及基于微调的范式。我们讨论了它们的实施细节、优点和针对文本到SQL的特定适应。

SQL任务从多个实现角度。

期望与未来方向：我们讨论了基于大型语言模型的文本到SQL的剩余挑战和限制，如现实世界的健壮性、计算效率、数据隐私以及扩展。我们还概述了潜在的未来研究方向和改进及优化的机会。

我们希望这项调查能为近期的研究提供一个清晰的概览，并激发未来的研究。图3展示了一个分类树，总结了我们调查的结构和内容。

II. O VERVIEW

文本到SQL是一个旨在将自然语言问题转换成可在关系数据库中执行的相应SQL查询的任务。正式来说，给定一个用户问题 Q（也被称为用户查询、自然语言问题等）