# （一）基于LLM的NL2SQL系统

## 系统背景:

近年来，随着互联网的飞速发展，各类数据呈爆炸式增长，其中蕴含了价值丰富的知识和信息，能够支持许多智能分析应用的需求，而这些海量高价值数据通常以数据库形式进行存储。结构化查询语言SQL是访问和处理结构化数据的重要工具。但传统的数据库信息查询手段所面临的问题日益凸显，主要表现在以下几个方面：

(1) 结构化查询语言门槛高

(2) 专业人员依赖性强，普通用户自主分析数据困难

(3) 分析周期长，依靠人工进行数据转换和查询语句编写

(4) SQL语句与用户思维方式存在脱节，表达业务逻辑不直观

(5) 对数据库的依赖性强，通用性和可拓展性较差

鉴于这种现状，研究将自然语言直接映射到SQL查询的NL2SQL技术，具有重要的理论价值和应用前景。近年来，依托大规模预训练语言模型LLM的深度学习技术为NL2SQL任务带来了全新的契机。本系统拟基于LLM，研究构建端到端的NL2SQL系统，以支持用户通过自然语言查询来访问和分析结构化数据，实现知识的深度挖掘。

## 主要功能

NL2SQL系统开发的主要目的是结合大规模预训练语言模型，将自然语言直接映射到SQL查询，以便捷地与数据库交互，主要功能模块包括指令Prompt构建、偏差校准和SQL生成三大模块。

（1）指令Prompt构建模块

支持用户通过自然语言对需求进行查询表达，根据该问题，召回与查询相关的数据库表及字段信息，构建包含数据库schema信息的指令prompt。

（2）偏差校准模块

构建提示以校准大语言模型的固有偏差，以历史对话形式注入，指导大语言模型生成更准确的SQL语句。

（3）SQL生成模块

针对每个问题，生成多个推理路径和SQL，将生成的SQL语句提交到数据库执行，返回最一致的查询结果。

## 开发环境

操作系统及版本：MacOS

支撑软件及版本列表：Pycharm + python 3.10

数据库环境：Sqlite

## 基本的架构体系

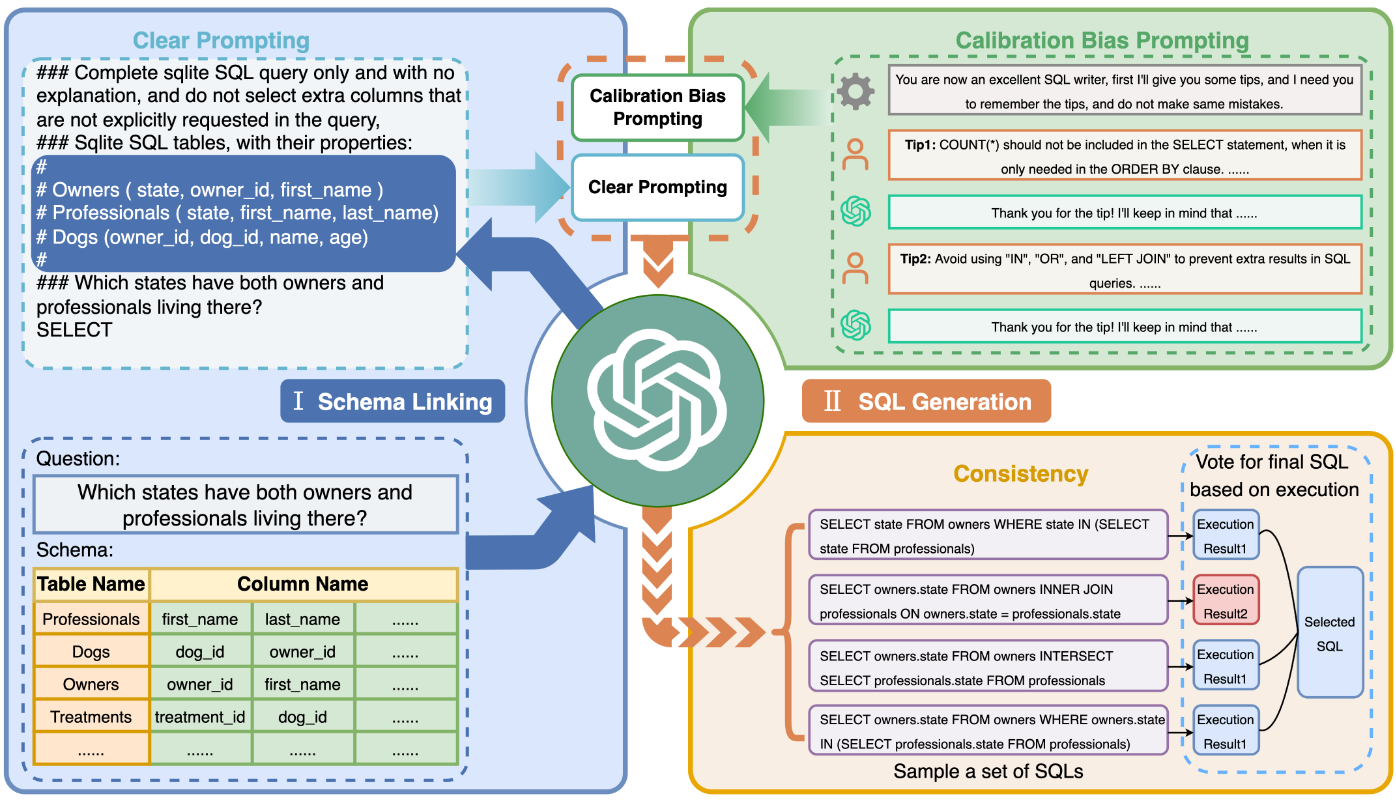


图 1 NL2SQL系统功能结构图

（1）指令Prompt构建模块

用户输入的自然语言问题，模型基于问题召回相关的数据库表及字段信息，最终拼接设计好的指令，共同构建一种结构化的零样本NL2SQL提示。

（2）偏差校准模块

此模块为一种插件式的校准策略，首先设定大语言模型的角色为优秀的SQL专家，接下来针对大语言模型的固有偏差分别构建提示，并以历史对话形式注入，以指导大语言模型生成更符合预期的SQL查询。

（3）SQL生成模块

对于每个问题，首先抽样多个推理路径以生成多样化的SQL回答，在数据库上执行并收集执行结果。在移除所有结果中的错误后，应用投票机制，选择最一致的查询结果及对应的SQL，并将结果返回给用户。