# 北航数据挖掘课程Project

基于大模型的自然语言转数据库查询语句研究

## 背景介绍

关系数据库能够存储大量信息，为医患信息记录、金融投资市场和客户关系管理等方面提供了技术基础。但是，访问关系数据库需要了解 SQL 等查询语言，仍具有一定的学习门槛[1]。近年来，大型语言模型（LLMs）在自然语言处理领域展现出了强大的能力，这些模型不仅能够理解和生成人类语言，还能够执行复杂的推理任务。其中，自然语言转SQL查询语句（Natural Language to SQL, NL2SQL）转变了人们与数据库交互的方式。

传统数据库查询需要用户掌握SQL语言，这对非技术背景的用户构成了较高的使用门槛。NL2SQL技术允许用户使用日常语言提问，系统自动将其转换为结构化的SQL查询语句，从而大幅降低了数据访问的技术门槛。

WikiSQL[2]是一个语义解析数据集，专注于单表查询场景，包含自然语言问题和对应的SQL查询语句，数据涵盖来自维基百科的不同表格，覆盖了广泛的领域知识，并且每个问题都对应一个相对简单的SQL查询，不包含表连接、嵌套查询等复杂操作，是评估和改进大模型NL2SQL任务能力的理想数据集。

近年来，随着Claude、Llama等大型语言模型的出现，NL2SQL技术取得了显著进步，这些模型通过大规模预训练和指令微调，展现出了理解自然语言指令并生成准确SQL查询的能力。在这一背景下，本项目旨在基于WikiSQL数据集的子集，探索大型语言模型在单表NL2SQL任务上的能力，评估模型的性能，并尝试通过提示工程等方法提升模型效果。

## 项目目标

1. **数据预处理与分析**
   1. 数据预处理：WikiSQL数据集中给出了训练集、验证机和测试集，训练集包含5万多个不同的问题-SQL对，验证集包含了8千多个问题-SQL对，请选择合适的训练集和验证集比例，合理抽样，确保数据分布的代表性。
   2. 数据集分析：对所选子集进行分析，包括问题类型分布、SQL查询复杂度、表格特征等。对数据集中的SQL查询进行分类，如简单选择查询、条件过滤查询、聚合查询等，并分析各类型的分布情况。
2. **大模型在WikiSQL上的表现评估**
   1. 模型选择：选择大语言模型进行评估，可以是：1个开源模型（如Llama2、Baichuan、ChatGLM等）或是商业API模型（如GPT-3.5、Claude等）
   2. 基线测试：设计基础提示模板，评估所选模型在无特殊优化情况下的WikiSQL转换能力。
3. **针对WikiSQL的提示工程优化**
   1. 提示词策略优化：设计1-2种不同的提示工程策略，例如示例学习以及思维链等，引导模型正确理解问题和表格。
   2. 跨查询类型分析：评估提示策略在不同类型查询上的效果差异。
4. **模型性能评估与分析**
   1. 评估指标：

执行准确率（Execution Accuracy）：生成的SQL能够被成功执行（SQL语法正确）

逻辑正确率（Logical Accuracy）: 生成的SQL执行结果与正确答案匹配的比例

## 数据格式

本项目的数据集包含 \*.jsonl 、 \*.tables.jsonl 两类后缀文件，每一类后缀都分为训练集（train）、测试集（test）和验证集（dev），因此一共有6个数据文件，均存放在 dataset 文件夹下。模型应在训练集和验证集上进行训练及训练期间的评估，并在测试集上进行指标验证。

\*.jsonl 类文件包含了问题和对应的SQL查询。数据集的每行是一条独立的查询以及涉及的表格操作等，包含4个字段，其中一行如下所示：

图形用户界面, 文本

AI 生成的内容可能不正确。

图1 \*.jsonl 类单行数据示例

这些字段含义如下：

|  |  |
| --- | --- |
| 字段名 | 说明 |
| phase | 数据收集阶段（WikiSQL由两个阶段收集完成） |
| question | 人工编写的提问 |
| table\_id | 提问对应的表格的ID |
| sql | 回答提问对应的 SQL 查询。它包含以下子字段:   * sel : 查询语句（select）中的列索引 * agg: 聚合操作类型（0:无, 1:MAX, 2:MIN, 3:COUNT, 4:SUM, 5:AVG） * conds：三元组列表，代表条件语句（where）列表，依次为 [列索引, 操作符索引, 条件值]，操作符索引：0:==, 1:>, 2:<, 3:>=, 4:<=，条件的比较值，为string或float类型。 |

表 1 \*.jsonl数据集字段说明

\*.tables.jsonl 类文件则记录了各类数据集涉及的表格原始信息。每一行都是一个完整的表格，包含5个字段。其中一行如下图所示：

图表, 散点图

AI 生成的内容可能不正确。

图2 \*.tables.jsonl 类单行数据示例

这些字段含义如下：

id：表格ID。

header：表格各列列名。

types：表格各列内容表示变量的类型。

rows：表格各行的值，每一列是表格的一整行。

Names：表格表头名。

## 注意事项

1. 本项目的数据集可以直接从[WikiSQL GitHub 项目](https://github.com/salesforce/WikiSQL/tree/master)下载，具体数据获取方法和评测方法请参考该项目的 README.md 文件，其中提供了详细的数据说明和评测脚本。在使用测评脚本时，请注意评测脚本仅接收JSONL格式的查询语句。
2. 【提交格式】各小组结课时需提供项目的实现代码、实验报告、答辩PPT，提交格式如图 2，提交链接<https://bhpan.buaa.edu.cn/link/AACEC41FE1795344C98248B155C2136702>
3. 实验报告按照《北航数据挖掘课程实验报告模板-2025春》模板撰写，链接<https://bhpan.buaa.edu.cn/link/AA434EA470CD5F44DB8A21B68AEB01DE7D> 请注意两个大作业分开撰写和提交。

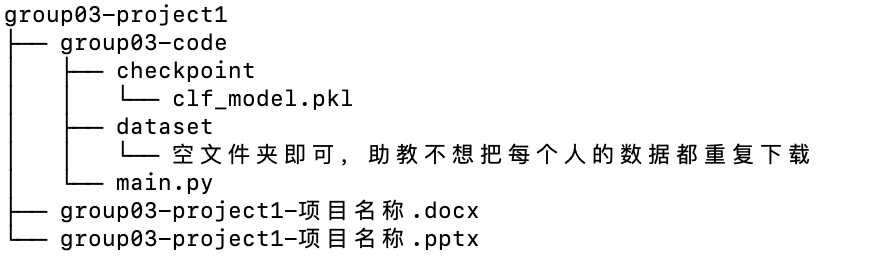


图 2 提交格式示例

1. 【代码规范】建议使用python语言完成项目。项目代码应具备结果可复现性，结果不能复现的、与报告汇报结果差距较大且不能合理说明理由的，按抄袭处理。
2. 【checkpoint】如果模型训练耗时较长，可以修改代码逻辑为"如果模型checkpoint已经存在，则直接加载模型，否则重新训练"，方便助教评测。
3. 【数据预处理】为避免在数据预处理过程中误操作导致的数据泄露，建议代码包含数据预处理算法，方便助教检查或有可能复现从原始数据到模型训练的全流程。
4. 项目代码和报告严禁抄袭，发现抄袭且不能合理说明理由的，一律按0分处理。
5. 项目报告严禁使用ChatGPT等AIGC写作工具直接生成[2]，发现明显AI写作痕迹的，酌情扣分。
6. 本文档将随课程进度更新，以最新版为准。更新通知会发到课程群。

编辑：陈一文

更新日期：2025年3月1日

## 参考文献

1. https://arxiv.org/abs/1709.00103
2. 本项目数据来自https://github.com/salesforce/WikiSQL/tree/master