🧪 5003 Group Project Report of LLMs4SQL

1. 小组信息

- **Siyu Xie (72542207) 本科专业:食品科学与工程**
- **Xinfang Zhang (72542152) 本科专业:信息管理与信息系统**
- **Jingyi Dong (72542072) 本科专业:大数据管理与应用**
- **Wenyue Yang (72542268) 本科专业:信息管理与信息系统**
- **Jingwen Luo (72542176) 本科专业:信息与计算科学**

2. 贡献说明

- **Siyu Xie (72542207) :**
  - 大模型云平台接口代码封装
  - `missing_token` 评估结果分析 + 报告
- **Xinfang Zhang (72542152) :**
  - 模型推理代码pipeline封装
  - `syntax_error` 评估结果分析 + 报告
- **Jingyi Dong (72542072) :**
  - 论文搜索 + 论文分析
  - `query_performance` 评估结果分析 + 报告
- **Wenyue Yang (72542268) :**
  - 结果评估代码pipeline封装
  - `query_equality` 评估结果分析 + 报告
- **Jingwen Luo (72542176) :**
  - 数据预处理代码 + 数据映射代码
  - 报告整合撰写 + 代码仓库整合

注：小组成员均进行了**原论文的研读分析+优化创新方案**的设计思考，且在项目过程中进行了**不同模块的代码的实现**。

3. 项目概述

3.1 背景 & 动机

近年来大型语言模型（LLMs）在自然语言处理、代码生成等领域表现强劲，但它们是否能“真正理解”结构化查询语言（SQL）仍存在疑问。结构化查询语言对语法、语义、上下文、执行逻辑都有严格要求。原论文提出，通过一系列 SQL-centric 任务全面评估 LLMs 的“理解能力”。

3.2 核心论文

- **论文名称**：*Evaluating SQL Understanding in Large Language Models* (arXiv)
- **作者**：Ananya Rahaman, Anny Zheng, Mostafa Milani, Fei Chiang, Rachel Pottinger (arXiv)
- **发表 / 提交时间**：2024 年 10 月 (arXiv)；后为 EDBT 2025 会议录 (卷 28, pp. 909-921) (experts.mcmaster.ca)

- **研究目标**：
  评估 LLMs 在 SQL 任务上的表现，考察它们在识别 (recognition)、上下文 (context)、语义 (semantics)、连贯性 (coherence) 这四个能力维度上的强弱。论文设计了一系列任务，包括语法错误检测 (syntax error detection)、缺失 token 识别 (missing token identification)、查询性能预测 (query performance prediction)、查询等价性检查 (query equivalence checking)、以及 SQL → 自然语言解释 (query explanation) 。
- **主要发现**
  尽管某些模型（如 GPT-4）在基础任务（recognition / context）上表现较好，但所有模型在更深层的语义理解与连贯性 (尤其是等价性判断与性能估计) 上仍存在显著不足。也就是说，目前 LLMs 在“真正理解复杂 SQL 语义与逻辑”的能力上仍有明显局限。

3.3 复现的动机 & 目标

- **复现动机**
  i. 原论文虽提供了任务定义与评估思路，但缺乏完整的可复现 pipeline —— prompt 设计、输出解析、评估流程等没有公开。
  ii. 实际部署与科研复现中，需要一个结构化、工程化、可扩展的框架，以便不同模型 / 不同 prompt / 不同任务的统一评估。
  iii. 通过系统化重构与扩展 (任务维度更细、输出结构化、可自动评估)，构建一个 “可信赖的 SQL 理解能力基准 (benchmark)” —— 对未来研究与系统应用都有现实价值。
- **复现目标**
  i. 基于论文任务定义，对语法错误检测、缺失 token 识别、查询性能预测、查询等价性分析等任务进行重构扩展。

ii. 设计结构化 prompt + JSON schema 输出 + 稳定 sampling + 统一 LLM server 接口。
iii. 构建 inference pipeline + evaluation pipeline，使评估结果可自动计算 (二分类、多分类、位置预测、F1 / MAE / Hit-Rate 等)。
iv. 支持多种 LLM 的 plug-and-play 测试。
v. 由于原论文的结果可复现性存疑，本项目只基于原论文的任务定义和数据，自行开发复现pipeline进行对比分析，不依赖原论文的实验结果。

4. 技术设计

4.1 论文方法概述

原论文提出了基于 SQL 的五类任务（语法错误检测、缺失 token 识别、查询性能预测、查询等价性检查、查询解释），核心技术思路如下：

1. **任务定义**：每类任务明确输入（SQL 查询）与输出（错误标记、性能类别、等价性标签等）。
2. **评估指标**：主要使用 Precision、Recall、F1-Score 等统计指标，部分任务引入 MAE、Hit Rate 等衡量模型预测的精确性。
3. **Prompt 使用**：论文通过简单自然语言提示指导模型生成答案，但未明确输出结构或 JSON schema，也未统一随机性采样策略。
4. **数据与实验**：数据来源包括 SDSS、SQLShare、Join-Order 和 Spider；实验结果部分经过人工评估，存在复现性与自动化不足问题。

4.2 本项目实现策略

针对原论文方法的局限性，本项目设计了完整、可复现的技术实现框架，包括：

1. **数据处理与标准化**
   - 收集与清洗原论文数据集，统一字段、表结构信息。
   - 对查询进行 tokenization、结构化标注，方便后续任务解析与定位。
2. **Prompt Engineering**
   - 设计 zero-shot 与 few-shot prompt，确保模型输出符合 JSON schema，便于自动解析。
   - 引入角色、工作流程约束，引导模型生成连贯、可解析的回答。
3. **Inference Pipeline**
   - 实现 SQL-centric 任务的推理 pipeline，支持不同模型 / 不同 prompt / 不同任务的统一评估。
   - 构建统一 LLM 接口（Doubao / Qwen / Deepseek / GLM 等），支持思考 / 非思考模式。
   - 支持批量推理、随机性控制与可复现采样。
4. **Evaluation Pipeline**
   - 对不同任务进行二分类、多分类、位置预测等评估。

- 自动计算 Precision / Recall / F1 / MAE / Hit Rate 等指标，生成可视化结果表格与分析报告。

5. **结果分析**
   - 提供任务维度的模型表现分析，发现模型在复杂语义和等价性任务上存在局限。
   - 不与原论文实验结果对比，以确保数据可靠性和实验可复现性。

4.3 与论文方法的偏差说明

- 本项目不直接使用论文的人工评估结果，而是通过统一、自动化的 pipeline 重新计算指标。
- 由于原论文未提供具体模型版本、prompt 或采样细节，本项目的实验环境与论文略有差异，因此不与其数值做直接对比。
- 选择以下模型进行实验对比分析
  - **Doubao-Seed-1.6-251015** ~ 来源：字节跳动最新可控思考模型
  - **qwen3-next-80b-a3b-instruct** ~ 来源：阿里巴巴 / 通义千问 开源系列最新非思考模型，MOE架构
  - **GLM-4.6** ~ 来源：智谱AI最新可控思考模型
  - **DeepSeek-V3.1-Terminus** ~ 来源：深度求索DeepSeek推出的可控思考模型
  - **DeepSeek-V3.1-Terminus (开启推理)** ~ 来源：深度求索DeepSeek推出的可控思考模型
- 输出格式、评估流程、任务划分在本项目中进行了优化，以提高可复现性与工程化水平。

5.算法 / 系统实现（Algorithm / System Implementation）

5.1 核心算法描述

本项目构建了一个**统一的 LLM SQL 理解评估系统**，核心设计如下：

1. **多平台 LLM 接口封装**
   - `LLMServer` 提供统一接口，支持 Doubao、Qwen、SiliconFlow、普通 OpenAI 接口等多平台。
   - 支持 `chat` 、 `vision chat` 、 `embedding` 模式，模型是否具有“推理能力”可配置。
   - 后端统一封装 API 请求，保证推理流程可复现。
2. **推理 Pipeline (`Inference` 类)**
   - 输入：SQL 查询 + 任务类型 (`InferType`)
   - 输出：符合 JSON schema 的结构化结果，便于自动化解析。
   - 支持批量处理与多线程推理（ `max_workers` 控制并行数量）。
   - 可配置推理策略：是否启用 reasoning / thinking 模式。
3. **评估 Pipeline (`EvaluateTool` 类)**

- 对不同任务类型进行自动评估，包括二分类、多分类、位置预测等。
- 指标支持：Precision / Recall / F1 / MAE / Hit Rate。
- 支持 Macro F1 计算，用于多分类任务的全局表现衡量。
- 自动加载评估数据，输出可视化结果与分析报告。

4. **配置驱动设计**
   - 使用 YAML 配置文件统一管理模型参数、推理策略、评估任务及数据路径。
   - 支持灵活切换模型和任务，无需修改核心代码。

示例配置（`config.yaml`）：

```yaml
model:
  model_type: doubao
  base_url: https://ark.cn-beijing.volces.com/api/v3/
  api_key: api_key
  reasoning_ability: True

inference:
  model_name: model_name
  model_identifier: NULL
  reasoning: False
  max_workers: 10

evaluation:
  infer_type: syntax_error
  data_dir: outputs/syntax_error
  model_list: ['DeepSeek-V3.1-Terminus-Thinking', 'GLM-4.6', 'DeepSeek-V3.1-Terminus',
```

5.2 关键数据结构

- **`Inference` 类**
  - `llms: LLMServer` — LLM 接口对象
  - `infer_type: InferType` — 推理任务类型
  - `model_name / model_identifier` — 模型标识
  - `max_workers` — 并行推理线程数
- **`EvaluateTool` 类**
  - `dataset` — 对应任务的数据集
  - `infer_type` — 评估任务类型
  - `metrics` — 自动计算 Precision / Recall / F1 / MAE / Hit Rate
- **推理输出**

```
{
  "syntax_error": "YES/NO",
  "syntax_error_type": <type>
}
```

5.3 正确性验证方法

- **数据一致性检查**：对输入 SQL 查询与标注数据进行 token / schema 对齐，保证推理输入有效。
- **结果解析校验**：对 JSON 输出进行严格 schema 校验，避免解析错误导致指标偏差。
- **指标对比**：对多分类任务，使用 Macro F1 全局衡量，确保不同模型可直接比较。
- **多模型 & 多任务复测**：使用多线程并行验证，确保评估结果的稳定性与可复现性。