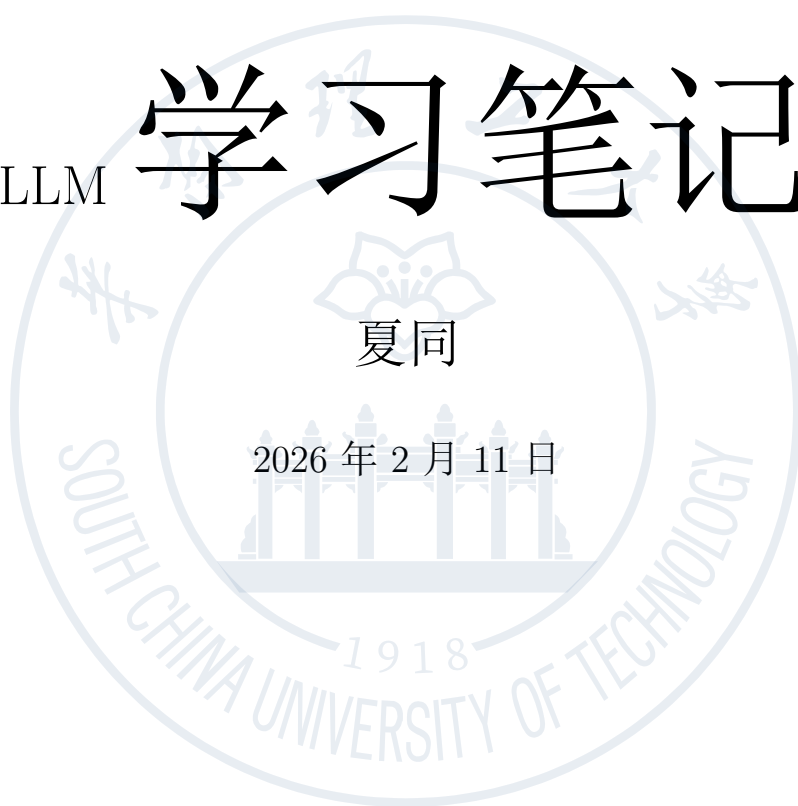


LLM

# 学习笔记

夏同

2026 年 2 月 11 日



# 目录

第一章	SQLENS: 文本到 SQL 语义错误检测与纠正的端到端框架	2
1.0.1	核心概念定义	2
1.0.2	总结	3
1.1	引入	3
1.1.1	目标	4
1.1.2	相关工作	4
1.2	方法论	4
1.2.1	问题定义	4
1.2.2	SQLENS 错误检测	4
1.2.3	信号聚合	5
1.2.4	SQLENS 错误纠正	5
1.3	实验评估	6
1.3.1	实验设置	6
1.3.2	主要结果	6
1.3.3	消融研究	7
1.3.4	个体信号的有效性	7
1.4	结论	7
1.5	局限性	8

# 第一章 SQUELS: 文本到 SQL 语义错误检测与纠正的端到端框架

文本到 SQL 系统 (text-to-SQL system) 将自然语言问题转换为 SQL 查询, 使得非技术用户能够与结构化数据交互。虽然大语言模型在文本到 SQL 任务上显示出有希望的结果, 但它们经常产生语义错误但语法有效的查询, 且对其可靠性缺乏深入理解。

该论文提出 SQUELS, 一个用于细粒度检测和纠正 LLM 生成 SQL 中语义错误的端到端框架。SQUELS 集成来自底层数据库和 LLM 的错误信号来识别 SQL 子句中的潜在语义错误, 并进一步利用这些信号指导查询纠正。在两个公开基准测试上的实证结果表明, SQUELS 在错误检测的 F1 分数上比最佳基于 LLM 的自评估方法高出 25.78%, 并将现成文本到 SQL 系统的执行准确率提升高达 20%。

## 1.0.1 核心概念定义

### 细粒度检测

在文本到 SQL 任务的上下文中, **细粒度检测**特指对大型语言模型生成的 SQL 查询进行深入的结构化分析, 旨在精准地定位和分类语义错误的精确位置与具体类型。与仅判断查询整体正确性的粗粒度评估不同, 细粒度检测具备以下特征:

- **分析粒度:** 以 SQL 查询的子句为基本分析单元 (如 SELECT、FROM、WHERE、GROUP BY、JOIN ON 等), 而非整个查询语句。
- **错误类型:** 专门针对**语义错误**, 即语法有效但逻辑与用户意图或数据库现实状态不符的查询。常见错误类型包括:
  - **模式引用错误:** 引用不存在的表、列, 或错误拼写。
  - **值域不匹配错误:** 谓词条件中的值不在目标列的值分布内 (例如, 查询 “city = 北京” 但该列仅有 “上海”、“广州” )。
  - **聚合误用错误:** 错误地使用 (或遗漏) 聚合函数 (如 COUNT, SUM)。
  - **连接条件错误:** 多表连接时关联条件不正确, 导致笛卡尔积或丢失关联。
- **输出形式:** 产生结构化的错误信号, 明确指出错误所在的子句及其错误类别, 而非简单的二分类 (正确/错误) 结果。

**比喻:** 细粒度检测如同一位经验丰富的数据库管理员进行代码审查, 不仅指出 SQL 无法

运行, 更精确标注出“WHERE 子句中的列名拼写错误”或“JOIN 条件遗漏了关键关联字段”。

## 端到端框架

在 SQUELS 的上下文中, 端到端框架指一个将细粒度错误检测与基于反馈的查询纠正两个核心阶段无缝集成的自动化系统。其“端到端”特性体现在:

- **输入与输出:**
  - **输入端:** 原始自然语言问题  $Q$  和初始 LLM 生成的、可能包含错误的 SQL 查询  $S_{initial}$ 。
  - **输出端:** 经过检测与纠正后, 语义正确性显著提升的最终 SQL 查询  $S_{corrected}$ 。
- **内部工作流:** 框架内部形成一个自动化、信息闭环的处理流水线, 无需人工干预:
  1. **统一信号采集:** 同时利用底层数据库的元信息、值分布、约束, 以及 LLM 自身的推理能力, 对  $S_{initial}$  进行细粒度分析, 生成结构化错误诊断报告  $\mathcal{E}$ 。
  2. **闭环引导式纠正:** 将错误报告  $\mathcal{E}$  作为明确的、结构化的指导信号, 输入到纠正模块 (通常仍为 LLM), 引导其针对性地修正  $S_{initial}$  中已识别的具体错误, 生成  $S_{corrected}$ 。
- **关键特征:** 两个阶段共享内部状态与信号, 纠正阶段直接利用检测阶段输出的精细化诊断信息, 而非独立运作。

**比喻:** 该框架如同一条智能装配线, 视觉检测单元 (细粒度检测) 发现特定螺丝未拧紧, 立即将“位置 A, 螺丝松动”的指令发送给机械臂 (纠正模块), 机械臂精准地拧紧该螺丝, 完成产品修正。整个过程自动连续, 信息无损传递。

### 1.0.2 总结

在 SQUELS 框架中:

- **细粒度检测**代表了其分析的深度与精度, 使其具备“精准诊断”的能力。
- **端到端框架**代表了其系统架构的集成度与自动化水平, 实现了从“诊断”到“治疗”的连贯、自主闭环。

二者的结合, 使 SQUELS 不仅能更准确地定位 LLM 生成 SQL 中的深层语义错误, 还能高效地利用这些诊断信息驱动查询的自动纠正, 从而系统性提升文本到 SQL 系统的可靠性。

## 1.1 引入

文本到 SQL 系统通过将自然语言问题转换为 SQL 查询, 使非技术用户能够查询关系数据库并提取见解。大语言模型显著推进了这一任务, 但 LLM 生成的查询仍然容易出错。文本到 SQL 系统的主要差距是缺乏细粒度、可解释的错误检测, 特别是对于语义错误 (查询执行但返回错误结果)。

### 1.1.1 目标

- 挑战 1: 在子句级别识别语义错误
- 挑战 2: 使用噪声错误信号预测 SQL 查询的正确性
- 挑战 3: 在没有正确性预言的情况下修复 SQL 查询

### 1.1.2 相关工作

#### 文本到 SQL

基于 LLM 的文本到 SQL 系统（如 DIN-SQL、MAC-SQL）使用多智能体框架将任务分解为子任务。其他方法如 MCS-SQL、CHESS 和 Chase-SQL 遵循类似的工作流程。最近还探索了基于强化学习的文本到 SQL 推理模型。

#### 文本到 SQL 中的错误检测与纠正

早期方法主要关注二元正确性分类，提供有限的错误性质洞察。基于 LLM 的文本到 SQL 系统通常依赖 SQL 执行反馈或 LLM 自反思来判断正确性。相比之下，SQUEL 产生可解释的子句级错误信号，可以指导用户调试和下游学习框架。

## 1.2 方法论

### 1.2.1 问题定义

- 定义 1（文本到 SQL 算法）：算法  $f$  接受自然语言问题  $Q$ 、数据库实例  $D$  和可选的外部知识  $K$ ，生成 SQL 查询  $q = f(Q, D, K)$
- 定义 2（语义错误）：语义错误  $e$  导致 SQL 查询  $q$  无法正确回答自然语言查询  $Q$ ，形式化为  $do(e) \Rightarrow O(q, D) \neq O(Q, D)$

### 1.2.2 SQUEL 错误检测

SQUEL 的错误检测器结合来自数据库和 LLM 的信号来检测语义错误。常见的语义错误类别包括：

#### 数据库错误信号

- 异常结果信号：检测 SQL 输出是否异常
- 空谓词信号：检测 SQL 查询中是否产生空结果的谓词
- 子查询中错误过滤信号：检测子查询中的问题过滤条件
- 错误 GROUP BY 信号：检测错误使用的 GROUP BY 子句
- 错误连接谓词信号：检测无效的连接谓词

- 次优连接树信号: 检测是否使用最优连接树
- 表相似性信号: 检测潜在的表选择错误
- 不必要子查询信号: 检测过度使用子查询
- 值歧义信号: 检测值使用中的歧义

### LLM 错误信号

- 列歧义信号: 识别数据库中与 SQL 查询中使用列相似的列
- 证据违反信号: 识别生成的 SQL 查询是否违反问题中指定的证据
- 证据不足信号: 评估可用证据是否足以确认 SQL 查询正确回答问题
- 错误问题子句链接信号: 评估 LLM 对生成 SQL 子句的信心
- LLM 自检信号: LLM 对自身输出的整体评估

### 1.2.3 信号聚合

SQUEL 采用弱监督框架聚合这些噪声信号, 将每个信号视为提供部分和噪声监督的标记函数。使用生成模型估计联合分布  $p(\Lambda, Y)$ , 其中  $Y$  是未观察到的真实标签。

### 1.2.4 SQUEL 错误纠正

SQUEL 的错误纠正策略采用分解方法, 逐步指导 LLM 修复查询:

#### 错误报告

每个错误信号与包含以下内容的详细错误报告关联:

- 信号描述
- 修正指令
- 识别为潜在错误源的问题子句
- 基于错误检测权重的置信度分级

#### 纠正组件

- 错误选择器: 使用 LLM 确定首先修复哪个错误
- 错误修复器: 应用针对性修正并使用 SQL 解析器验证修订后的查询
- SQL 审计器: 在原始版本和修订版本之间选择最佳查询
- 初始化步骤:
  - 调用 ErrorDetector 检测初始 SQL 查询  $q$  中的所有语义错误, 得到错误集合  $\mathcal{E}$
  - 初始化迭代计数器  $i \leftarrow 0$
  - 初始化修正后的查询  $q' \leftarrow q$
- 迭代修正循环:

- 循环条件: 错误集合  $\mathcal{E}$  非空且迭代次数  $i$  小于最大迭代次数  $max\_iter$
- 每次迭代执行:
  - \* 调用 ErrorSelector 从当前错误集合  $\mathcal{E}$  中选择优先级最高的错误  $e$
  - \* 调用 ErrorFixer 修复错误  $e$ , 生成修正后的查询  $q'$
  - \* 从错误信号集合  $\mathcal{S}$  中移除已处理的错误信号  $e$
  - \* 调用 ErrorDetector 重新检测修正后查询  $q'$  中的错误, 更新错误集合  $\mathcal{E}$
- 护栏信号检查:
  - 检查护栏信号  $e_g$  是否触发 (返回集合大小为 1)
  - 如果触发, 调用 ErrorFixer 修复护栏信号对应的错误
- 最终审计与返回:
  - 调用 SQLAuditor 审计修正后的查询  $q'$  和原始查询  $q$
  - 返回最优的修正查询  $q'$

## 1.3 实验评估

### 1.3.1 实验设置

在 BIRD 和 Spider 两个标准文本到 SQL 基准测试的开发集上进行评估。使用四种最先进的文本到 SQL 系统生成 SQL 查询进行评估。

### 1.3.2 主要结果

SQUEL 在所有现成的文本到 SQL 系统上一致提高了端到端准确率, 并在 BIRD 上优于自反思和 Fix-ALL。在现实环境中, SQUEL 实现了最高的净改进, 修复了更多查询并引入了更少的回归。

表 1.1: BIRD 上的端到端准确率提升

方法	$\Delta Acc.(N_{net})$	$\Delta Acc.(N_{fix})$	$N_{net}$	$N_{fix}$	$N_{break}$
Vanilla (初始准确率: 59.07%)					
Self-Reflection	59.07(+0.00%)	60.6(+1.53%)	1	22	21
SQUEL w. Fix-ALL	61.15(+2.08%)	62.34(+3.27%)	30	47	17
SQUEL	62.54(+3.47%)	63.66(+4.59%)	50	66	16
DIN-SQL (初始准确率: 39.49%)					
Self-Reflection	54.63(+15.14%)	56.37(+16.88%)	209	233	24
SQUEL w. Fix-ALL	58.11(+18.62%)	59.71(+20.22%)	257	279	22
SQUEL	59.99(+20.50%)	61.45(+21.96%)	283	303	20

### 1.3.3 消融研究

#### 错误检测性能分析

SQUELS 在预测语义正确性方面优于所有基线方法。对于 DIN-SQL, SQUELS 实现了 78.88 的 F1 分数, 比最佳 LLM 自评估基线高出 21.45 点。

表 1.2: BIRD 上 SQUELS 错误检测效果

方法	准确率	AUC	精确率	召回率	F1
Vanilla					
LLM Self-Eval.(Bool)	60.53( $\pm 2.30$ )	X	57.70( $\pm 18.90$ )	10.02( $\pm 4.54$ )	16.97( $\pm 7.35$ )
SQUELS	64.63( $\pm 1.97$ )	61.90( $\pm 2.18$ )	58.11( $\pm 2.80$ )	48.74( $\pm 4.31$ )	52.94( $\pm 3.24$ )
DIN-SQL					
LLM Self-Eval.(Bool)	61.52( $\pm 1.20$ )	X	86.83( $\pm 2.18$ )	42.99( $\pm 2.72$ )	57.43( $\pm 2.20$ )
SQUELS	75.29( $\pm 2.33$ )	81.49( $\pm 1.74$ )	81.64( $\pm 2.17$ )	76.41( $\pm 3.50$ )	78.88( $\pm 2.19$ )

#### SQL 审计器和护栏信号的影响

SQL 审计器有助于减少修正过程中破坏的 SQL 查询数量, 但以修复较少查询为代价。护栏信号的使用通过增加修复总数同时减少回归总数来提高净改进。

### 1.3.4 个体信号的有效性

#### 错误检测中的个体信号

对于 MAC-SQL, 14 个信号中有 13 个超过 60% 的精确率。异常结果信号在识别 40 个错误时达到 100% 精确率, 次优连接树信号检测到最多错误 (62% 精确率)。

#### 错误纠正中的个体信号

在 MAC-SQL 生成的 SQL 查询上评估个体错误信号的纠正效果。在数据库信号中, 次优连接树信号影响最强, 空谓词信号也表现良好。在 LLM 信号中, 证据违反信号产生最高的净增益。

## 1.4 结论

我们提出了 SQUELS, 一个利用数据库和基于 LLM 的错误信号进行文本到 SQL 中子句级语义错误检测和纠正的新框架。SQUELS 使用弱监督聚合噪声信号, 预测查询正确性, 并

应用 LLM 指导的迭代纠正。它在错误检测方面优于 LLM 自评估，并在不依赖正确性预言的情况下将现成文本到 SQL 系统的执行准确率提升高达 20%。

## 1.5 局限性

SQUEL 在 SQL 错误检测和纠正中利用了 LLM，因此继承了底层 LLM 的某些限制。SQUEL 还产生了聚合多个信号和执行迭代纠正的计算开销，可能在实时环境中引入高延迟。然而，SQUEL 旨在作为离线的通用 SQL 调试工具，设计为异步细化和验证生成的查询。

