面向国产数据库的 Text-to-SQL 数据集设计

李国深[1]，刘莹君[2]，于莉娜[2]，纪　涛[2]，张　航[1]，吴继冰[1]

（1. 大数据与决策国家级重点实验室，湖南　长沙　410073；2. 智能空间信息国家级重点实验室，北京　100029）

摘　要：随着智能技术的发展，数据库数量和规模激增，传统数据存取技术在应对海量数据处理需求时存在耗时长、效率低等问题，Text-to-SQL 技术成为衔接用户需求和数据库存取的重要桥梁。然而，现有技术通常在开源非国产数据集上训练，在实际应用中存在数据库操作语言不一致、领域知识欠缺和可靠性差等问题。为此，结合数据库领域软硬件国产化趋势，设计面向国产数据库的 Text-to-SQL 数据集，采用基于合成数据方法的大语言模型两阶段训练技术，提出一种基于大语言模型的国产数据库 Text-to-SQL 方法，通过实验对方法的有效性进行了充分验证。

关键词：大语言模型微调；合成数据集；偏好学习；国产数据库

中图分类号：TP311. 138　　**文献标识码：**A　　**DOI：**10. 19358/j. issn. 2097 - 1788. 2025. 11. 009

引用格式：李国深，刘莹君，于莉娜，等. 面向国产数据库的 Text-to-SQL 数据集设计［J］. 网络安全与数据治理，2025，44(11)：52 - 59.

The design of Text-to-SQL datasets for domestic databases

Li Guoshen[1], Liu Yingjun[2], Yu Lina[2], Ji Tao[2], Zhang Hang[1], Wu Jibing[1]

(1. National Key Laboratory of Big Data and Decision, Changsha 410073, China;
2. National Key Laboratory of Intelligent Geospatial Information, Beijing 100029, China)

Abstract: With the development of intelligent technology, the number and scale of databases have surged. Traditional data access technologies face problems such as long-time consumption and low efficiency when meeting the needs of massive data processing. Text-to-SQL technology has thus become an important bridge connecting user needs and database access. However, existing technologies are usually trained on open-source non-domestic datasets, and their application is plagued by issues like inconsistent database operation languages, lack of domain knowledge, and poor reliability. To address this, this paper, in line with the localization trend of software and hardware in the database field, designs a Text-to-SQL dataset for domestic databases, adopts a two-stage training technology for large language models based on synthetic data methods, proposes a Text-to-SQL method for domestic databases based on large language models, and fully verifies the effectiveness of the method through experiments.

Key words: fine-tuning of large language models; synthetic dataset; preference learning; domestic database

0　引言

文本到结构化查询语言（Text-to-SQL，T2S）是自然语言问题和数据库工具结合的重要研究领域，具体是指将自然语言转化为计算机可执行的 SQL 查询语句的过程，它解决了从非结构化的自然语言和数据库模式到结构化 SQL 的转换等系列问题。T2S 技术的核心在于从文本数据里自动识别专业术语、所属领域、关联关系及结构特征，进而构建相应映射体系。传统映射构建模式高度依赖领域专家的人工规范操作，这种方式在知识体系持续迭代更新，或者领域专家资源匮乏的场景下，往往会暴露出耗时久、成本高、易出错等诸多弊端。而随着自然语言处理技术的迅猛发展，大语言模型与 T2S 技术的融合应用已成为新的发展趋势。

传统的 T2S 方法是基于规则模式的语法解析和模板匹配，需要大量人工标注或手动构建规则[1]。而大语言模型具有强大的语言理解和生成能力[2]，能够理解文本内容、提取关键信息、识别语义关系。利用大语言模型对大规模文本进行预训练，可从中自动学习实体和关系以及数据库模式，进而构建和更新从文本到 SQL 的映射关系，减轻领域专家在数据标注、规则构建阶段的工作量。然而，当前 Text-to-SQL 研究的进展仍受限于数据集的质量与规模[3]。现有主流数据集如 Spider、WikiSQL、

Bird 虽在多领域覆盖与复杂查询标注上取得一定成果，但仍存在领域分布不均衡、真实业务场景模拟不足、标注成本高昂等问题[4]，难以满足实际应用中多样化的 SQL 查询需求。与此同时，合成数据技术凭借其高效、低成本的优势展现出巨大潜力[5]，特别是训练数据数量匮乏条件下，在数据增强与模型泛化能力提升方面表现突出。

综上，本文采用国产达梦数据库（DM）开展数据集设计，达梦数据库作为国产数据库系统之一，在军事、政务等关键领域逐步替代 Oracle 等国外数据库。本文针对“执勤”业务场景，设计国产数据库系统并构建专用数据集，该数据集包含 300 条高质量标注样本，主要针对军事典型业务查询场景。达梦数据库的模式权限设计参考《达梦数据库技术文档》[6]。同时，采用基于合成数据方法的大语言模型两阶段训练技术，通过对比实验评估合成数据与真实数据的分布一致性及对模型性能的提升效果，探索大语言模型在国产数据库环境下的适配方法，为数据保障业务提供技术支撑。实验结果表明，本数据集不仅能有效补充现有数据资源的不足，且通过合成数据验证的方式，为 Text-to-SQL 数据集的构建与评估提供了新的技术路径。

1　国产数据库与数据集

1.1　国产数据库设计

1.1.1　系统概述

本系统涉及的单位实体有执勤人员、执勤装备、执勤任务、执勤地点、执勤安全等级，能够描述执勤任务信息，以及执勤任务规划、执勤单位描述、装备管理等任务。

1.1.2　需求描述与分析

本执勤信息系统旨在为数据保障单位提供全面、高效的管理解决方案，满足其在信息管理方面的需求。系统需具备以下功能：

（1）人员信息管理：系统应能记录执勤人员的基本信息，包括姓名、年龄、联系方式、在岗年限等，并支持信息的录入、修改、查询和删除，以便上级在分配任务或执勤时能够快速了解执勤人员的基本信息。

（2）领导信息管理：系统应能记录单位领导的基本信息，包括姓名、职称、经历、干龄等，并支持信息的录入、修改、查询和删除，方便上级下达任务时在组织层面能够便捷地进行信息管理。

（3）装备信息管理：系统应能记录装备的基本信息，包括装备名称、规格、生产厂家、使用次数、使用年限、历次任务出现情况等，并支持信息的录入、修改、查询和删除。此外，还需能记录装备的库存进出情况，以便管理人员及时掌握库存信息。

（4）任务信息管理：系统应支持任务的输入输出，同时需提供任务信息查询功能，方便上级实时了解任务完成情况。

（5）背景信息管理：任务在下达后，系统应能记录当时环境情况和任务执行信息，保证后续能够根据任务数据和背景信息进行意图识别和自动任务分配等。

（6）业务目标：实现执勤任务分配、人员调度、地点监控、意图识别等核心功能，同时支持以自然语言形式查询实时数据，满足高效便捷的信息获取需求。

本研究使用的数据需要具备以下特点：

（1）高安全性：鉴于数据涉及人员隐私与任务机密，必须支持算法加密，例如 SM4 算法，以保障数据安全。

（2）实时性：执勤任务状态处于动态更新中，要求数据库具备高效的事务处理能力，确保数据及时准确反映任务实际情况。

（3）跨域关联：数据涵盖人员、装备、任务、地点等多个实体，实体间存在复杂的关联关系，需要支持跨域关联查询。

1.1.3　概念结构设计

借助 E-R 图（如图 1 所示）清晰定义核心实体及关系。数据库实体包括执勤人员（含 ID、姓名、岗位、联系方式、所属单位）、单位领导（含 ID、姓名、职级、联系方式、经历、所属单位）、执勤任务（含任务 ID、任务类型、开始时间、结束时间、任务性质、任务内容描述）、装备（含装备 ID、装备类型、参数、状态、数量、历次任务使用情况、所属单位）、执勤地点（含地点 ID、地点名称、经纬度、安全等级）。

关系则包括：

任务－人员：呈现多对多关系，即一个任务由多名人员执行，而一名人员可参与多个任务。

任务－领导：呈现多对多的关系，即一个任务由多名领导负责，而一名领导可负责多个任务。

领导－人员：呈现多对多的关系，即一个单位领导管理多名人员，一名人员受多个领导管理。

任务－装备：同样为多对多关系，一个任务需调配多种装备，一种装备也可用于多个任务。

任务－地点：为一对多关系，一个任务对应多个执勤地点。

针对达梦数据库的特点[6]，设计数据库时做以下处理：

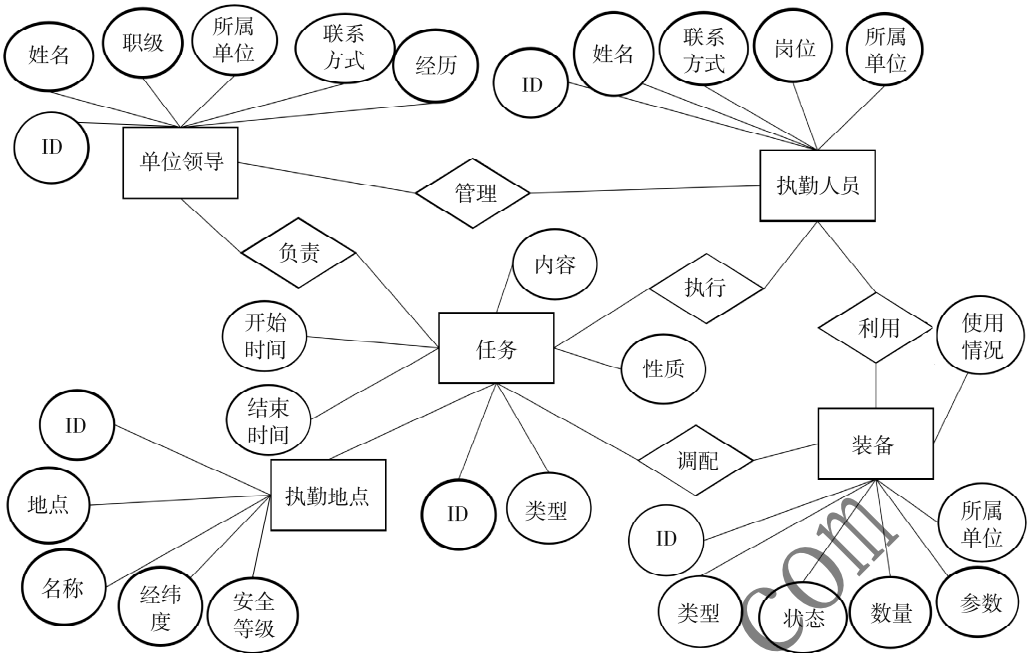

图 1　执勤数据库 E-R 图

索引组织表优化：对高频查询表（如执勤任务表 DM_TASK）采用聚簇索引，将任务 ID（TASK_ID）作为主键索引列，确保任务状态更新时的快速定位。

日期类型处理：任务时间字段（START_TIME，END_TIME）定义为 DATETIME 类型，并在数据库初始化时设置 COMPATIBLE_MODE = 2 以兼容 Oracle 的日期时间格式，避免时分秒信息丢失。

模式权限分离：创建 DUTY_QUERY 模式用于自然语言查询，与核心业务模式 DUTY_CORE 物理隔离，通过达梦的“用户 - 模式 - 角色”三级权限体系实现细粒度访问控制。

1.1.4　逻辑结构设计

在逻辑设计中，执勤人员表与任务表如表 1、表 2 所示，其余表设计类似，属性均在 E-R 图中展现。

表 1　执勤人员表（DM_PERSONNEL）

字段名	数据类型	约束条件	说明
PERSONNEL_ID	INT	主键，自增	人员唯一标识
NAME	VARCHAR（50）	非空	姓名
POSITION	VARCHAR（30）	非空	岗位
PHONE NUMBER	INT	非空	电话、联系方式
UNIT	VARCHAR（100）	非空	所属单位

根据数据库设计范式分析，所有表均满足第三范式（Three Normal Form，3NF），无隐含传递依赖。但是仍然存在隐藏风险：若业务中“岗位（POSITION）”与“所属单位（UNIT）”存在固定关联（如“干事”必属于“东路派出所”），则 UNIT 可能传递依赖于 POSITION，违反 3NF。但当前设计中 UNIT 直接依赖于 PERSONNEL_ ID，未暴露此问题（需根据实际业务规则验证）。

表 2　执勤任务表（DM_TASK）

字段名	数据类型	约束条件	说明
TASK_ID	INT	主键，自增	任务唯一标识
TASK_TYPE	VARCHAR（30）	非空	任务类型
START_TIME	TIMESTAMP	非空	任务开始时间
END_TIME	TIMESTAMP	可为空	任务结束时间
TASK_CONTENT	VARCHAR（30）	非空	任务内容
TASK_QUALITY	VARCHAR（30）	非空	任务性质

1.2　数据集构建

1.2.1　数据库系统构建

国产达梦数据库在功能、性能、使用场景、语句语法上与 Oracle 均有所不同[7]。下面介绍在数据保障任务中数据库系统的构建，如下伪代码所示：

```
-- 创建执勤数据库
CREATE DATABASE “DM_DUTY” DB_FILE_PATH '/dm/data/DM_DUTY. dbf' SIZE 1024;
USE DM_DUTY;
-- 创建表（示例：执勤人员表）
```

```
CREATE TABLE DM_PERSONNEL (
    "PERSONNEL" _ID INT PRIMARY KEY AUTO_INCREMENT,
    "NAME" VARCHAR (50) NOT NULL,
    "POSITION" VARCHAR (30) NOT NULL,
    "POSITION" VARCHAR (30) NOT NULL,
    "UNIT" VARCHAR (100) NOT NULL
); -- 其他表创建类似，略
```

在创建生成数据库和导入数据集时，根据达梦数据库“别名含特殊符号或长度超过 30 字节，必须加双引号”的特性，对所有的表名、列名两侧均加上“”。插入数据的格式如下所示：

```
-- 插入执勤人员
INSERT INTO "DM_PERSONNEL" (NAME, POSITION, UNIT) VALUES
('李明', '科长', '一中队'),
('王芳', '干事', '二中队');
--插入执勤任务
INSERT INTO "DM_TASK" (TASK_TYPE, START_TIME, DESCRIPTION) VALUES
('巡逻', '2025-05-06 08：00：00', '东门岗区域巡逻');
```

利用鱼台县综合行政执法局公布的开源数据集[8]，生成模拟执勤数据。在达梦数据库系统中导入存储，构建完整的数据库。开源数据集示例如表 3 所示。

表 3　模拟数据

带队领导	执勤人员	执勤中队	执勤时间	执勤地点	装备	任务
张军	杨慎	鱼台县执法大队	2021-06-18	鱼台	警棍	48421
张军	张鹏	鱼台县执法大队	2021-06-18	鱼台	警棍	48422
张军	丁荣	鱼台县执法大队	2021-06-18	鱼台	警棍	48423
张军	甄浩	鱼台县执法大队	2021-06-18	鱼台	警棍	48424

表 3 展示了执勤数据库中的任务数据信息，此外，其余分表对应了数据库结构中的其他表的详细数据，在此不再赘述。根据数据安全性要求，还需要对数据进行脱敏处理：采取模糊化人员姓名、任务细节，利用拼音替换等方式，使数据符合保密要求。

1.2.2　数据集导入

问题、数据库模式与标准的 SQL 语句是大模型 Text-to-SQL 问题的输入输出，针对执勤数据库，还需要设计执勤相关问题以及对应的标准 DM-SQL 查询才能获得完整的训练集。同 Spider 数据集[9]，对训练集中的问题以及 DM-SQL 查询进行难度分级，设计 300 条数据集，三种难度占比均为 33%，并从各难度取样例数据作为测试集。难度分级如表 4 所示。

表 4　SQL 难度分级

等级	判定标准	示例
简单	单表查询，仅含 SELECT（非聚合）+ WHERE（≤1 条件）	“查询气缸数 >4 的汽车数量” SELECT COUNT（*）FROM cars_data WHERE cylinders >4
中等	含 JOIN + GROUP BY（单字段），或 SELECT 聚合 + 多条件 WHERE	“按体育场统计音乐会数量” JOIN + GROUP BY 单字段（无嵌套）
困难	含 JOIN（≥3 表）+ GROUP BY（多字段）+ HAVING，或含 EXCEPT/UNION/单层嵌套	“欧洲有≥3 汽车制造商的国家” JOIN + GROUP BY + HAVING（无深层嵌套）

以下是训练数据示例：

```
{"db_id":"DM_DUTY",
" question":" Query the task type and task description of the task that uses the most equipment and has the longest task duration",
"SQL":"SELECT T. TASK_TYPE, T. DESCRIPTION FROM "DM_TASK" T JOIN "DM_TASK_EQUIPMENT" TE ON T. TASK_ID = TE. TASK_ID WHERE T. END_TIME-T. START_TIME = (SELECT MAX (END_TIME-STAR T_ TIME) FROM "DM_TASK") AND T. TASK_ID IN (SELECT TASK_ID FROM (SELECT TASK_ID, SUM (QUANTITY) AS total_quantity FROM "DM_TASK_EQUIPMENT" GROUP BY TASK_ID ORDER BY total_quantity DESC LIMIT 1));}
```

2　合成数据方法实验

设计执勤数据集后，本研究利用合成数据[5]的方法进行实验验证。

2.1　强数据生成

针对国产达梦数据库的语法特性与军事执勤业务场景，基于强模型 Qwen3-225b-a22b 采用“领域约束 - 语法适配 - 难度分级”的三层提示词设计框架生成强数据，强数据生成提示词示例如图 2 所示。具体步骤如下。

2.1.1　领域与语法适配的提示词设计

提示词需包含达梦数据库特有的语法规则[6]，例如：

表名与列名规范：所有表名/列名添加双引号（如"DM_TASK"），避免与关键字冲突。

函数适配：使用达梦兼容函数（如 TO_CHAR (START_TIME, 'YYYY-MM-DD') 替代 MySQL 的 DATE_FORMAT）。

事务控制：在复杂查询中加入 BEGIN WORK/COMMIT 语句，符合军事数据的强一致性需求。

```
Prompt for DM Data
Your task is to generate one additional data point at the {difficulty} difficulty level, in alignment with the format of the two
provided data points.
1. Domain: The domain can be expanded beyond the scope of "duty performance", to enhance the model's generalization ability
for the domestic database ecosystem.
2. Schema: Post your domain selection, craft an associated set of tables. These should feature logical columns, appropriate data
types, and clear relationships
3. Question Difficulty-{difficulty}:
  - Easy: Simple queries focusing on a single table.
  - Medium: More comprehensive queries involving joins or aggregate functions across multiple tables.
  - Hard: Complex queries demanding deep comprehension, with answers that use multiple advanced features.
4. Answer: Formulate the SQL query that accurately addresses your question and is syntactically correct.
  - SQL must contain the syntax structure corresponding to the difficulty
Additional Guidelines:
- Ensure SQL uses multiple tables for higher difficulty levels
- Maintain diversity in question topics
--Use functions and syntax compatible with Dameng Database, such as TO_DATE() to handle the time format, and the transaction
control statement BEGIN WORK/COMMIT. And add "" to both sides of all table names
Examples: {example_block}
Generate ONLY the following sections:
Domain: [Your Domain Here]
Schema:
CREATE TABLE table1 (...);
CREATE TABLE table2 (...);
Question:[Your Question Here]
Answer:[Your SQL Query Here]
DO NOT:
- Generate incorrect Question,answer and Schema
- The generated questions and sql queries exceed the The regulation of difficulty or do meet the difficulty requirements
- Add any additional comments,symbols
- A line wrap is required after a section
```

图 2　强数据生成提示词（prompt）

2.1.2　跨领域数据扩展与难度分级

除核心执勤场景外，生成气象数据库、地形数据库等关联领域数据，提升模型泛化能力。例如，在提示词中加入“生成涉及地形数据与执勤路线规划的跨领域查询”。

难度分级标准：

简单：单表查询（如 SELECT "NAME" FROM "DM_PERSONNEL" WHERE "POSITION" ='哨兵'）。

中等：多表 JOIN（如 SELECT "TASK_TYPE", "EQUIP_TYPE" FROM "DM_TASK" JOIN "DM_TASK_EQUIPMENT" ON ...）。

困难：嵌套子查询 + 聚合（如上述示例中的三层子查询）。

2.1.3　数据筛选与验证机制

语法验证：通过达梦数据库执行器校验生成的 SQL，过滤因引号缺失、函数误用导致的语法错误。

领域去重：排除教育、医疗等非军事领域数据，确保数据与业务场景强相关。

2.2　弱数据生成

利用弱模型（如 DeepSeek-Coder-1.3B）生成达梦数据库特有的错误样本，通过“错误注入 - 执行反馈 - 正负样本标注”构建偏好数据集。

2.2.1　达梦数据库典型错误模拟

语法错误：遗漏表名双引号（如 SELECT * FROM DM_TASK）、误用 MySQL 函数（如 DATE_FORMAT(START_TIME, '%Y-%m')）。

语义错误：跨表关联时忽略外键约束（如错误关联"DM_TASK" 与"DM_PERSONNEL" 表）。

事务缺失：复杂查询未添加 COMMIT，导致数据未持久化。

2.2.2　正负样本标注与数据集构建

正向验证：弱模型生成的 SQL 若在达梦数据库中执行结果与真实标签一致，标记为正样本 y_w。

反向验证：若执行报错或结果错误（如因表名缺失双引号导致语法错误），标记为负样本 y_l。

数据增强：对正样本进行同义词替换（如“执勤人员”→“哨兵”），对负样本注入不同类型错误（如替换 TO_CHAR 为 DATE_FORMAT），扩充数据集多样性。

2.3　监督微调

监督微调阶段利用生成的强数据以及基础数据集对基础开源模型 CodeLLaMa-7B 进行训练，采用交叉熵损失函数进行监督微调，优化模型生成符合输入提示的目标 SQL 语句 y 的条件概率分布，目标函数为：

$$E_{(x,y)} \sim D_s\left[\sum_{t=1}^{T} \log p_\theta\left(y_t \mid y_{1:t-1}, x\right)\right] \tag{1}$$

其中，D_s 为强数据集，x 为包含数据库模式和自然语言问题的输入提示，y 为目标 SQL 语句，θ 为模型参数。$p_\theta(y \mid x) = \prod_{t=1}^{T} p_\theta\left(y_t \mid y_{<t}, x\right)$ 是给定提示 x 的情况下目标 SQL 语句 y 的条件概率分布。T 是 y 的序列长度，t 是自回归解码步骤。训练过程中采用余弦学习率衰减策略，初始学习率设为 2×10^{-5}，批量大小为 16，训练 3 个 epoch，使模型初步掌握复杂 SQL 生成能力。

2.4　偏好学习

采用直接偏好优化（Direct Preference Optimization，DPO）算法基于偏好数据对经过强数据微调的模型进一步训练，目标是最大化模型对正样本的偏好得分，跳过奖励建模阶段，同时抑制负样本的生成概率。该阶段旨在最大化以下目标函数：

$$L_{\mathrm{dpd}} = E\left(x, y_w, y_l\right) \sim D_w \log\sigma\left(\beta\log\frac{p_\theta\left(y_w \mid x\right)}{p_{\mathrm{ref}}\left(y_w \mid x\right)} - \beta\log\frac{p_\theta\left(y_l \mid x\right)}{p_{\mathrm{ref}}\left(y_l \mid x\right)}\right) \tag{2}$$

其中，p_{ref} 为参考模型（如强数据微调后的初始模型）的概率分布，β 为调节参数（实验中设为 0.2），σ 为 Sigmoid 函数。该函数通过比较模型与参考模型在正负样本上的输出差异，引导模型学习执行器偏好。

实验细节上，偏好数据集采用 1 ∶ 1 的正负样本比例进行批量采样，避免类别不平衡影响训练效果，使用 Adam 优化器，学习率设为 2×10^{-6}，低于强数据微调阶段，以避免过度破坏已学习的强数据模式。同时，在输入提示中加入历史交互信息（如多轮对话中前序 SQL 语句），增强模型对上下文依赖的处理能力，适用于多轮 Text-to-SQL 场景。

3 实验结果及分析

3.1 实验环境与模型选择

实验环境配置如表 5 所示。实验在两张 NVIDIA A800GPU 上进行，采用 PyTorch2.5 框架和 DeepSpeed 加速库。训练过程中使用混合精度（BF16）和梯度裁剪（最大梯度范数为 1.0）以提升训练效率并避免梯度爆炸。

表 5 实验配置

	配置信息	参数
编译环境	PyTorch	2.5.1
语言	Python	3.12
GPU	A800-80 GB	×2
CPU	Intel Xeon Processor	14Cores
RAM	/	150 GB
基座模型	CodeLLaMa-7B	/

基础模型选择 CodeLLaMa-7B 作为主要模型，其具备较强的代码生成能力和逻辑推理能力，适合作为 Text-to-SQL 任务的起点。后续还可以在 CodeLLaMa-13B 上实现，以验证其性能。

弱模型则采用 DeepSeek-Coder-1.3B 作为弱模型生成模型，该模型参数规模较小，在 Text-to-SQL 任务中容易产生语法错误、表与列名错配等典型问题，能够有效生成正负样本，构建偏好数据集，让模型进行偏好学习。

强模型使用最新发布的 Qwen3-235b-a22b 作为强数据生成模型，利用其强大的跨域知识储备、推理能力以及生成能力，生成高质量的跨域 Text-to-SQL 数据集。

3.2 数据集

实验采用在基准数据集的基础上生成强弱数据的方式获得训练集，其包括了基准数据集以及合成数据集。

执勤数据集是模拟执勤场景设计的小规模面向军事领域的数据集，包含 1 个数据库、210 条训练集以及 97 条测试集。

强数据生成：通过 Qwen3-235b-a22b 大语言模型生成合成数据，提示词模板遵循“Domain-Schema-Question-Answer”四元组结构，由于提前设置了领域多样性控制、难度分级和格式规范，确保生成数据与基准数据集兼容。最终生成 500 条强数据，覆盖多种场景。

弱数据生成：使用 DeepSeek-Coder-1.3B 在执勤数据库上生成 SQL 语句，通过 SQL 执行器验证，执行结果与真实标签一致的为正样本，不一致的为负样本，构建包含 210 条正负样本的数据集。然而，在偏好学习阶段，对于每一条 question，其应该都有正负两个样例作为训练输入，所以在弱数据生成阶段得到的正样本还需要通过数据扰动变成负样本，以便于所有的训练数据 question 都有正负样例输入。这样做的目的是保证模型能够得到更全面的错误模式。数据扰动模式包括表名替换、列名替换、运算符反转、删除关键条件、破坏子查询以及最简单的随机删除关键词等。

3.3 难度差异对比

表 6 展示了 T2S 模型与其他模型在执勤数据集不同难度等级测试集上的性能对比：相较于目前主流的推理大模型，TS2 的参数量远远小于 DeepSeek-r1 与 Qwen-3-238b-a22b，但是在 Text-to-SQL 领域的性能达到了几乎相同的地步，这将有利于将 T2S 广泛部署使用，也验证了此方法可有效降低模型对参数量的依赖，为资源受限场景提供可行方案。

表 6 模型性能对比

模型	简单/%	中等/%	困难/%	总和/%
DeepSeek-r1	87.88	64.1	32	63.92
Qwen-3-238b-a22b	84.4	61.5	24	59.8
T2S	74.3	60.6	24	56.7

此外，T2S 模型在简单领域的准确率达到了 74.3%，但是在复杂领域的准确率不高，根本原因是导入的社会基准数据集与生成的强数据不足。T2S 模型对于模型领域泛化和复杂推理能力的提高作用不显著，而针对于单表的简单 SQL 查询在强数据中有充分生成，使得模型在简单问题中仍然能够保持优势。

利用 API 调用强模型 DeepSeek-v3 对此数据集进行测试，其结果同样在中等和困难问题领域的准确性不高。通过调查发现，这类如“执勤”数据集具有机密性、专业性，故决定了其专有数据难以直接用于公开模型，所以在这类数据集上的表现效果不佳，从侧面也印证了数据的重要性。

3.4 消融实验

根据表 7 消融实验结果，强模型消融移除强数据后（w/o Strong），T2S 在执行准确率上从 56.7% 显著下降至 45.3%，这种下滑表明强数据凭借其更简单的 SQL 查询以及对领域通用化的强调，能为模型性能提供有力支撑，是提升准确性的关键支柱。同时，当强数据这个支柱移除后，弱数据的作用便体现出来：禁用偏好学习后即无弱数据（w/o Weak），T2S 在执勤数据集上的性能也有所下降，生成的 SQL 语句的语法错误增加。这说明，在强数据的支撑中断时，弱数据虽无法完全替代强数据的核心作用，却能通过抑制错误语法、减少典型错误的工作机制，为模型性能提供一定的兜底提升，如图 3 所示。其维持模型基本运转、降低错误率的功能是切实有效的。

表 7　消融实验

模型	执行准确率/%
T2S	56. 7
w/o Weak	53. 6
w/o Strong	45. 3

注："w/o Weak"表示无弱数据（without Weak）；"w/o Strong"表示无强数据（without Strong）。

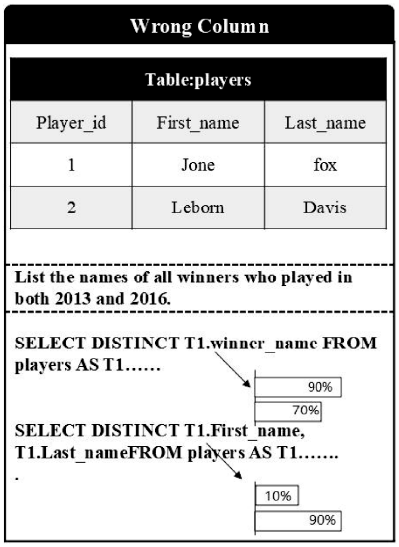

图 3　语法错误对比（偏好学习作用）

3. 5　参数量影响分析

通过观察调用开源 API 对执勤数据集进行测试的结果（如表 8 所示）可知，随着参数量的增加，模型在 Text-to-SQL 领域的效果随之提高。根据图 4 不难看出，T2S 模型总准确率均高于开源模型 Qwen1. 5；并且所有模型在困难领域的准确率均较低。同时，Qwen-1. 5-72B 的激活参数多于 T2S，故其训练成本相对较高[10]，但是总体准确率却低于 T2S，突显 T2S 的训练成本优势。此外，Qwen1. 5-14B 模型的性能远远高于 Qwen-1. 5-7B、Qwen-1. 5-0. 5B，而与 Qwen-1. 5-72B 模型性能相差不大，这说明随着参数量提高，模型性能也能随之提升，但是存在瓶颈。因此，与同参数量的模型以及小模型对比后发现 T2S 的参数量既满足效率，也没有浪费额外的模型参数，说明其具备参数量优势。

表 8　不同参数模型性能对比

模型	简单/%	中等/%	困难/%	总和/%
Qwen-1. 5-0. 5B	3. 03	2. 56	0	2. 06
Qwen-1. 5-7B	60. 6	43. 6	4	39. 2
Qwen-1. 5-14B	75. 7	53. 8	12	50. 5
Qwen-1. 5-72B	75. 76	61. 5	12	53. 6
T2S	74. 3	60. 6	24	56. 7

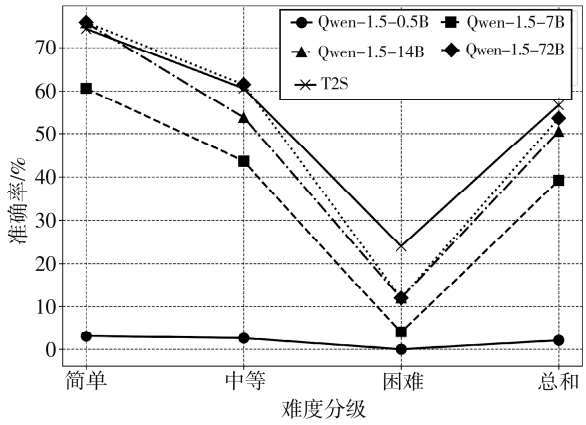

图 4　不同参数模型性能对比

3. 6　模型影响实验

调用 API 测试推理和非推理模型在执勤数据集上的性能，结果如表 9 所示，其中，非推理模型 DeepSeek-v3 性能表现更优异，输出上更加规范，且推理模型的生成序列较长，思维链（Chain-of-Thought，CoT）步数多于非推理模型[11]，其时间开销明显较高。

表 9　不同参数模型性能对比

模型	简单/%	中等/%	困难/%	总和/%
DeepSeek-r1	87. 88	64. 1	32	63. 92
DeepSeek-v3	93. 4	66. 6	24	64. 9

3. 7　实验结论

本实验主要围绕国产数据库达梦，通过将大模型融入达梦数据库的语法特性和业务场景，结合强弱数据生成技术，有效提升了模型对国产数据库的适配能力。实验结果表明，本研究所构建的数据集能够显著提高模型在国产化环境下的查询准确率和语法合规率，为大语言模型在国产数据库领域的应用奠定了坚实基础。

4　结论

本文围绕国产数据库达梦，详细阐述了 Text-to-SQL 数据集的设计与构建过程，包括数据库概念结构、逻辑设计、数据集生成方法及实验验证。通过融入达梦数据库的语法特性和业务场景，结合强弱数据生成技术，以及根据预试验和正式实验的结果，得出以下结论：首先，针对国产达梦数据库构建的专用 Text-to-SQL 数据集，成功填补了国产数据库适配 Text-to-SQL 技术的数据集空白，融合强数据语法适配与弱数据错误验证的构建逻辑，为

同类国产化场景数据集的开发提供了可复用范式；其次，该数据集与大语言模型两阶段训练技术的结合，有效突破了现有技术在国产数据库环境下语法不兼容、领域知识不足的核心瓶颈，显著提升了模型在执勤等关键业务场景的实用性与可靠性，为 Text-to-SQL 技术在国产化软硬件生态中的深度落地奠定了核心数据与方法支撑。最终模型在达梦数据库上的语法合规率和业务场景准确率均显著优于基准模型，为国产数据库的自然语言交互提供了可行方案。然而受限于军事数据敏感性，当前数据集仅基于公开政务数据模拟，真实执勤场景的复杂约束（如跨单位任务协同、多级安全审批）尚未完全覆盖，未来需结合脱敏后的真实数据进一步验证。

未来可进一步关注困难领域准确率提升、语法差异与安全约束，探索动态提示优化技术，进一步提升模型对达梦特有函数[7]（如 WM_CONCAT）的支持能力以及将扩展数据集覆盖更多国产数据库（如人大金仓、神舟通用），还可以探索“DM 数据库 + 国产大模型（如 ChatGLM-6B）”的国产化方案，通过联邦学习技术或本地部署的方式[12]解决隐私数据训练问题，推动 Tent-to-SQL 在边防、仓储等场景的落地，推动 Text-to-SQL 技术在国产化生态中的深度应用。

参考文献

[1] LI F, JAGADISH H V. Constructing an interactive natural language interface for relational databases [J]. Proceedings of the VLDB Endowment, 2014, 8 (1): 73 - 84.

[2] RADFORD A, NARASIMHAN K, SALIMANS T, et al. Improving language understanding by generative pre-training [EB/OL]. [2025 - 07 - 13]. openai. com/research/lan.

[3] QIN B, HUI B Y, WANG L H, et al. A survey on Text-to-SQL parsing: concepts, methods, and future directions [EB/OL]. (2022 - 08 - 29) [2025 - 08 - 06]. https://arxiv.org/abs/2208.13629.

[4] 邢朝军. 基于检索增强生成的文本转 SQL 优化和应用的研究 [D]. 上海：华东师范大学，2024.

[5] YANG J X, HUI B Y, YANG M, et al. Synthesizing Text-to-SQL data from weak and strong LLMs [C]//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Bangkok, Thailand: Association for Computational Linguistics, 2024: 7864 - 7875.

[6] 武汉达梦数据库有限公司. 达梦数据库管理系统 DM8 [Z]. 武汉：武汉达梦数据库有限公司，2019.

[7] 达梦技术社区. 浅谈达梦数据库与 MySQL 数据库的差异 [EB/OL]. (2023 - 08 - 04) [2025 - 7 - 16]. https://eco.dameng.com/community/article/c7aa5e1fb676238d037ba654732aaebe.

[8] 鱼台县综合行政执法局. 鱼台县综合行政执法局_龙虾节执勤信息 [EB/OL]. (2021 - 10 - 15) [2025 - 08 - 06]. https://data.sd.gov.cn/portal/catalog/9ca661dc84ac4b0db15fd6b42543ec9e.

[9] YU T, ZHANG R, YANG K, et al. Spider: a large-scale human-labeled dataset for complex and cross-domain semantic parsing and Text-to-SQL task [C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: Association for Computational Linguistics, 2018: 3911 - 3921.

[10] Qwen Team. Introducing Qwen1.5 [EB/OL]. (2024 - 02 - 04) [2025 - 08 - 6]. https://qwenlm.github.io/zh/blog/qwen1.5/.

[11] HASHEMI M, BAMGBOSE O, MADHUSUDHAN S T, et al. DNR Bench: benchmarking over-reasoning in reasoning LLMs [EB/OL]. (2025 - 03 - 20) [2025 - 08 - 06]. https://arxiv.org/abs/2503.15793.

[12] MCMAHAN H B, MOORE E, RAMAGE D, et al. Communication-efficient learning of deep networks from decentralized data [C]//Proceedings of the International Conference on Artificial Intelligence and Statistics. Cadiz, Spain: PMLR, 2016: 1273 - 1282.

（收稿日期：2025 - 07 - 24）

作者简介：

李国深（2003 - ），男，硕士研究生，主要研究方向：数据驱动智能决策、大语言模型。

刘莹君（1987 - ），女，本科，助理工程师，主要研究方向：大数据分析与决策。

吴继冰（1987 - ），通信作者，男，博士，副研究员，硕士生导师，主要研究方向：数据治理、大数据分析与决策。E-mail：wujibing@nudt.edu.cn。

版权声明

凡《网络安全与数据治理》录用的文章，如作者没有关于汇编权、翻译权、印刷权及电子版的复制权、信息网络传播权与发行权等版权的特殊声明，即视作该文章署名作者同意将该文章的汇编权、翻译权、印刷权及电子版的复制权、信息网络传播权与发行权授予本刊，本刊有权授权本刊合作数据库、合作媒体等合作伙伴使用。同时，本刊支付的稿酬已包含上述使用的费用，特此声明。

《网络安全与数据治理》编辑部