

# 智能问数应用

SQLCoder 构建大模型数据分析助手

李晓晨

2023.11.23

# 目录

**01** 智能问数场景介绍

**02** SQLCoder系列模型

**03** SQLCoder使用技巧详解

**04** SQLCoder应用实战演示

# 智能问数场景介绍

# 智能问数 (SQLCoder) 应用样板间

The screenshot displays the SQLCoder application interface. At the top, there is a header with the Baidu logo and the text 'SQLCoder-7B 模型API地址: kc-...jinal'. Below the header, there are navigation tabs for '数据管理' and '功能反馈'. The main content area shows a search result for the query '森林覆盖率最高的国家是哪些?'. The result is presented as a line chart with the X-axis labeled 'country' and the Y-axis labeled 'max\_forest\_coverage'. The chart shows a line representing the maximum forest coverage for various countries, with a tooltip for '苏里南' (Suriname) showing a value of 98.5. The chart also includes a legend for 'max\_forest\_coverage' and a '数据表格' (Data Table) tab. The interface also features a sidebar with '推荐提问' (Recommended Questions) and '历史提问' (History Questions) sections, and a footer with '优选专家服务 技术专家1V1定制服务' and a version number '0.0.1'.

## 离线部署说明

应用仅作为演示使用，请勿用于生产环境

Docker Windows Linux Mac 云上部署

### 1 下载应用

AMD64版本 ARM64版本(推荐)

### 2 解压文件，运行应用

解压文件后可得到 code-sql 文件，运行文件需要有ak/sk参数；  
获取ak/sk: <https://console.bce.baidu.com/qianfan/ais/console/applicationConsole/application> ;  
运行: 您可以双击 code-sql 运行，或者通过如下命令在终端运行:

```
/code-sql -ak xxxxx -sk xxxxx
```

注意: 如果想替换默认 8086 端口, 可通过如下命令在终端运行:

```
/code-sql -ak xxxxx -sk xxxxx -port 80
```

- 无缝兼容干帆平台API能力
- 支持本地下载，离线部署
- 干帆专属增值服务，协助企业开展深度定制

体验地址: <https://console.bce.baidu.com/tools/?u=bce-head#/sampleAppCenter>

## 领域属性

只遵循固定的问答模板，只回答sql，无需其他通用能力，但需要大模型对sql语法有强专业能力

Sql场景的prompt容易超长

Sql场景的评估难

## 业务相关

一个数据库对应的是一套业务，业务逻辑隐含在各个表的schema定义中，没有显示给出

表之间的结构复杂，长表、宽表；星型模型、雪花模型、星座模型

业务的额外要求（查询效率、字段使用、领域黑话）

## 业务架构

对接数据库的工程

操作数据库有风险，需要安全校验

可视化、用户交互体验

# SQLCoder系列模型



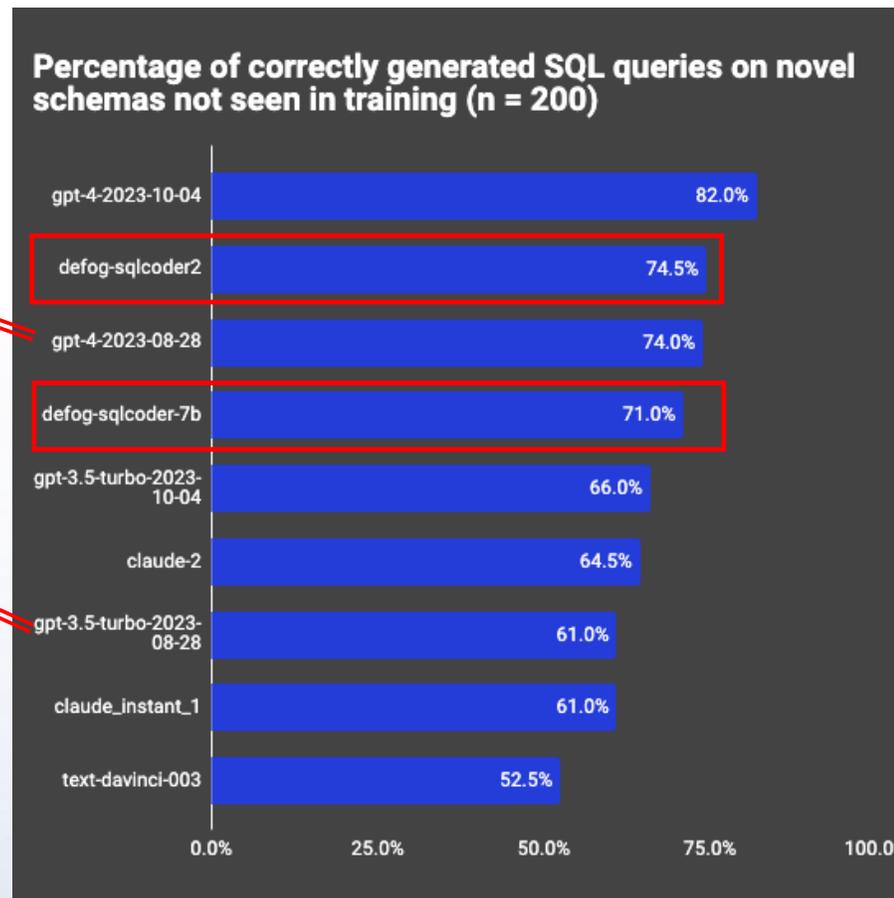
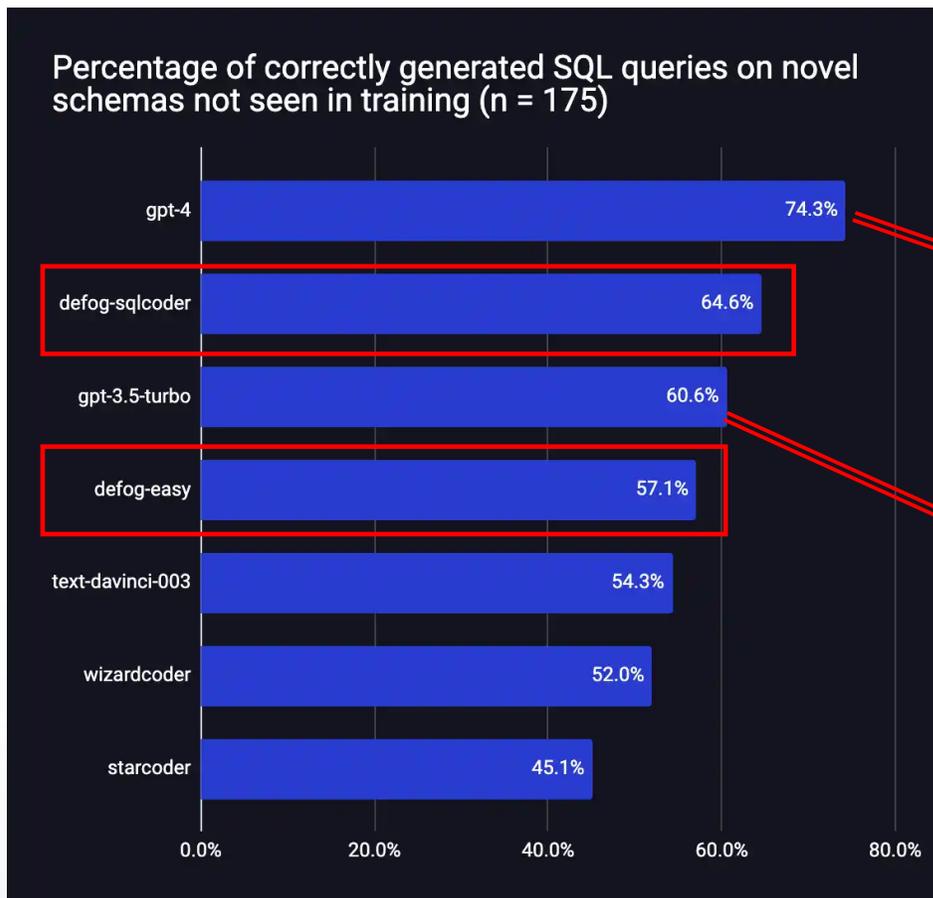
# Defog.ai SQLCoder family



	sqlcoder	sqlcoder2	sqlcoder-7b
发布日期	2023.08	2023.10	2023.10
模型参数量	15B	15B	7B
基础模型	StarCoder 15.5B <a href="https://huggingface.co/bigcode/starcoder">https://huggingface.co/bigcode/starcoder</a> 用1T token的代码数据上预训练的模型 支持8k上下文		Mistral-7B <a href="https://huggingface.co/mistralai/Mistral-7B-v0.1">https://huggingface.co/mistralai/Mistral-7B-v0.1</a> 代码能力约等于Code-Llama 7B, 保留了更强的泛化能力 支持8k上下文
训练数据	10,537 条人工标注的 text-to-SQL 问题 含有10种不同的schema	20,000条人工标注的 text-to-SQL 问题 含有10种不同的schema	
训练过程	数据被训练两个epoch。 问题被分了四种不同难度, 先训练简单和中等难度的数据, 再训练困难和极难的数据	在sqlcoder的基础上, 在训练数据配比、微调更新参数量、超参数选择 等方面做了优化	



# Defog.ai SQLCoder family





# Defog.ai SQLCoder family

### Accuracy by category on out-of-training-set schemas

	date	group_by	order_by	ratio	join	where
gpt-4	72	91.4	82.9	80	82.9	80
sqlcoder2-15b	76	80	77.1	60	77.1	77.1
sqlcoder-7b	64	82.9	74.3	54.3	74.3	74.3
gpt-3.5	68	77.1	68.6	37.1	71.4	74.3
claude-2	52	71.4	74.3	57.1	65.7	62.9
claude-instant	48	71.4	74.3	45.7	62.9	60
gpt-3	32	71.4	68.6	25.7	57.1	54.3

# SQLCoder 的局限性



## 英文较强，中文能力一般

- 预训练模型见过中文，但sql精调阶段都是英文数据
- 解决方法：可以进一步做中文sft



## 具体业务逻辑判断问题

- 通用sql模型，对具体的业务了解不深
- 解决方法：业务数据sft



## 没有chat能力

- 没有训练过对话数据，只有写sql的能力
- 解决方法：不使用此模型做chat场景，仅作单轮使用



## 主要遵从Postgre语法

- 精调阶段使用的是postgre语法
- 解决方法：
  - 在prompt中强调当前应该使用的语法
  - 使用对应语言的数据继续精调
- Sql方言翻译工具

# SQLCoder使用技巧详解

# 智能问数参考架构



# Prompt模板

```
### Task
Generate a SQL query to answer the following question:
`{question}`

### Database Schema
This query will run on a database whose schema is
represented in this string:

CREATE TABLE product_suppliers (
supplier_id INTEGER PRIMARY KEY, -- Unique ID for each
supplier
product_id INTEGER, -- Product ID supplied
supply_price DECIMAL(10,2) -- Unit price charged by
supplier
);
【其它表略】
-- sales.product_id can be joined with products.product_id
-- sales.customer_id can be joined with
customers.customer_id
-- sales.salesperson_id can be joined with
salespeople.salesperson_id
-- product_suppliers.product_id can be joined with
products.product_id

### SQL
Given the database schema, here is the SQL query that
answers `{question}`:
```sql
```

```
### 任务
生成一个SQL查询以回答以下问题:
`{question}`

### 数据库模式
此查询将在一个{sql type}数据库上运行 该数据库的模式在以下字符串中表示:

CREATE TABLE product_suppliers (
supplier_id INTEGER PRIMARY KEY, -- 每个供应商的唯一ID
product_id INTEGER, -- 供应的产品ID
supply_price DECIMAL(10,2) -- 供应商的单价
);
【其它表略】
-- sales.product_id 可以与 products.product_id 进行连接
-- sales.customer_id 可以与 customers.customer_id 进行连接
-- sales.salesperson_id 可以与 salespeople.salesperson_id 进行连接
-- product_suppliers.product_id 可以与 products.product_id 进行连接
{additional_comments}
{fewshot}
### SQL
根据数据库模式, 以下是回答 `{question}` 的SQL查询:
```sql
```

## 数据量

100条

~

万条

百条量级

这个量级优先保证数据的高质量，一般直接使用业务数据，使sqlcoder熟悉业务schema和使用习惯

千条量级

真实业务数据和自动扩充数据混合，增强泛化能力

万条量级

可以加入通用text2sql数据，全面提升sqlcoder的能力，注入新知识

## 数据获取

人工标注

构造Schema+自然语言问题+sql语句

大模型扩充

已有标注数据sql不变，生成问法不同、意义相同的自然语言问题

大模型标注

使用真实业务sql和schema，利用大模型生成对应的自然语言问题

## 数据飞轮

应用上线后收集真实query和反馈反哺训练数据

线上query

正确性反馈



增量sft

调整数据配比



训练方法

- 全量SFT: 更新全量参数, 微调效果好
- LoRA: 更新参数量少, 训练速度更快, 减少遗忘
- 首选全量SFT, LoRA适合特殊场景



超参配置

- SQL场景建议epoch不超过2, 减少过拟合现象
- 学习率一般使用推荐值0.00002即可



训练阶段

- 数据量较大时, 可以分批训练, 先易后难



增量迭代

- 增量数据可以在之前模型基础上训练
- 重新训练时, 数据配比倾向于实际错误率高的类型

# Sql场景的评估



人工打分



裁判员模型评估 (GPT-4/EB4)



数据库执行——通过率 / 正确率



开源评估框架 (spider、WikiSQL)



ROUGE/BLEU



AST parse



# SQLCoder应用实战演示

# 课后作业

## 基础作业:

1、基于课程示例数据集复现SQLCoder 调优过程，使用SQLCoder预置模型体验不同的Prompt并将模型测试效果、感受发布至社区

## 进阶作业:

1、使用SQLCoder 搭建自己专属的智能问数应用，对接数据库，体验10+条Prompt并将应用搭建经验发布至社区

**\*将上述作业在【百度智能云干帆社区】进行发布，发布时选择 #大模型实训营 话题**



扫码进入  
百度智能云干帆社区



添加小助手  
进入课程群

**Q & A**

**T H A N K S**