# CCKS2024 Archer: Bilingual Text-to-SQL evaluation

# 1、Introduction

Natural language interaction with databases in a more friendly and intuitive way is a challenging work, which aims to translate natural language questions into executable SQL statements. Some recent works have achieved good performance on existing datasets, but they cannot efficiently perform complex reasoning such as mathematics, common sense, and hypothesis. To this end, we propose Archer, a dataset that incorporates the above three types of inference to make more complex and subtle queries. In addition, we tested with both large language models and fine-tuned models. Even methods that achieve SOTA on existing datasets still only achieve 6.73% executable rate on our dataset, indicating that Archer is still a challenging dataset for current models and techniques.

Archer has three types of reasoning: mathematical reasoning, commonsense reasoning and hypothetical reasoning. Mathematical reasoning has an important proportion in the specific application scenarios of SQL. Commonsense reasoning refers to the ability to reason based on implicit commonsense knowledge, Archer contains some questions that require understanding the database to infer missing details; Hypothesis reasoning requires the model to have counterfactual thinking ability, which is the ability to imagine and reason about unseen situations based on visible facts and counterfactual hypotheses. Examples of the three types of reasoning are shown in the following figure:



# 2、Data description

Archer consists of 1042 Chinese questions, 1042 English questions, and 521 corresponding SQL queries, covering 20 different databases in 20 domains. Eight databases are used as the training set, two databases are used as the validation set and ten databases are used as the test set. Data set and the leaderboard address: https://sig4kg.github.io/archer-bench/

# 3、Evaluation metrics

We use two metrics to evaluate the ability of the model: VAlid SQL (VA) and EXecution accuracy (EX). VA is the fraction of predicted SQL statements that execute successfully, regardless of whether the answer is correct or not. EX is the fraction of predicted SQL statement execution results that match the standard SQL statement execution results.

# 4、Schedule of matches

This competition is in the form of Leaderboard, and the specific arrangements and requirements are as follows:

To submit your method to the Archer benchmark for evaluation, follow the instructions below and send the required files to archer.benchmark@gmail.com. Ensure that all submissions are thoroughly tested on the development set. Similar to the Bird benchmark, we accept three types of submissions:

**1) Up to 4 A100 80G GPU inference:**

Due to resource constraints, we support up to 4 A100 80G GPU inference to evaluate your method. (This limit may change over time).

**Requirements:**

Provide a detailed README file and compressed code (including requirements.txt).

The inferred cost estimate for the specified evaluation is about 5 times larger than your evaluation on the development set.

For example :1 A100 80G, 3 hours.

**Optional:**

Package the environment into a Docker image for easy download and evaluation.

Push your model to huggingface with the appropriate privacy rules.

 **2) Api-based inference:**

If your approach is based on an API-based LLM service (e.g., GPT family, Claude 2, etc.), follow this guidance. Please note that depending on the LLM service you are using, we may need API keys for evaluation.

**Requirements:**

Provide a detailed README file and compressed code (including requirements.txt).

The inferred cost estimate for the specified evaluation is about 5 times larger than your evaluation on the development set.

For example: $10, gpt-4-turbo.

**Optional:**

Package the environment into a Docker image for easy download and evaluation.

Push your model to huggingface with the appropriate privacy rules.

**3). Joint inference:**

If your approach involves both API-based inference and open source LLMs inference using Gpus, follow this combined inference submission guide. Similar to GPU inference, we support up to 4 A100 80G GPU inference to evaluate your method. (This limit may change over time). Depending on the LLM service you're using, we may need an API key for evaluation.

**Requirements:**

Provide a detailed README file and compressed code (including requirements.txt).

The inferred cost estimate for the specified evaluation is about 5 times larger than your evaluation on the development set.

Example: $10, gpt-4-turbo + 1 A100 80G, 3 hours

**Optional:**

Package the environment into a Docker image for easy download and evaluation.

Push your model to huggingface with the appropriate privacy rules.

Note that we allowed each team to provide up to 2 checkpoints and 3-4 submissions during the evaluation time. In addition, a maximum of 2 preferred results can be selected for updating per commit.

Sign up for a team at —————— August 20th - August 30th

Commit results and code —————— September 10th

Evaluation paper —————— September 15

Awards: —————— September 19-22

 Registration: (Email registration, need to send: captain, team member, team name, instructor, organization, contact information and captain email and other relevant information, **registration address: zhichaoyan@foxmail.com**)

Participating groups: Please scan the following QR code on wechat to join the group. If the QR code is invalid, please contact us to get the latest QR code.

**5. Rules**

1). Participants should ensure that the information submitted during registration is accurate and valid. All the competition qualification and bonus payment are subject to the information submitted;

2). The team shall be formed by email. The number of team members shall not exceed 3 (including the team leader), and the team list shall not be changed after the registration deadline.

3). Each team should appoint a team leader, and the team name should not exceed 15 characters. The team name should not violate Chinese laws and regulations or public order and good custom words, otherwise the organizer may disband the team;

4). Each player can only join one team. Once a player is found to join multiple teams by registering multiple accounts, the relevant team will be disqualified;

5). Allow the use of open source code or tools, but not any code or tools that are not publicly released or require license;

6. In addition to the data set provided by the organizers, participants are allowed to use publicly available pre-training data (such as word vectors, character vectors, etc.), and are allowed to use public data sets (such as Spider,CSpider, etc.) to pre-train or fine-tune large models to improve knowledge extraction capabilities (training data should be provided to the organizers for review in the code review stage!). It is strictly forbidden to label the test set or construct similar data for training.

**6：Rewards**

The total bonus pool of this task is RMB 10,000, and the bonus is set as follows:

First place: RMB 5000, 1 place

Second place: RMB 3000, 1 place

Third place: RMB 1000, 2 places

**7、Task organizers and contacts (email)**

**Organizers：**

Jeff Pan, University of Edinburgh（j.z.pan@ed.ac.uk）

Zhichao Yan，Shanxi University（zhichaoyan@foxmail.com）

Wenyu Huang, University of Edinburgh（）

**Academic Guidance Group:**

Jeff Pan, University of Edinburgh

Ru Li，Shanxi University

Mirela Lapata, University of Edinburgh

**Contacts：** Zhichao Yan（zhichaoyan@foxmail.com）

**8. Disclaimer:**

We will only use your code for evaluation purposes and will not disseminate or disclose any details of your code. We will remove your code and Docker immediately after the evaluation is complete and confirmed with the model authors.