



AT&T



H2O GenAI WORLD NEW YORK



Text-to-SQL:

Fine Tuning for Enhanced Query Generation

Farbod Tavakkoli – AT&T CDO – Data Scientist

Table of Contents

Introduction

Why Fine Tuning

Overview

Methodology

Fine Tuning

Data Curation

Data Profiling

Synthetic Data Generation

Model Card

H2O LLM Studio

Results

Introduction



Why Fine Tuning

Scaling laws: as compute spent in training goes ↗↗, model performance keeps improving

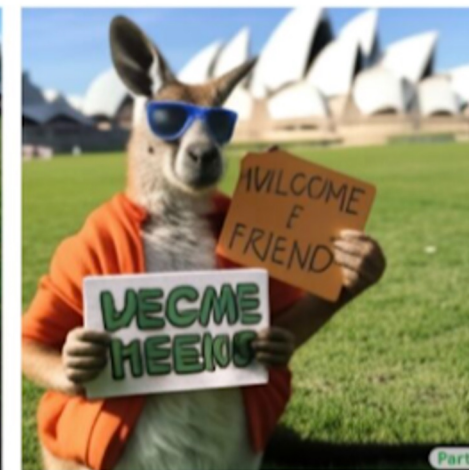
Parti-350M



Parti-750M



Parti-3B



Parti-20B



A portrait photo of a kangaroo wearing an orange hoodie and blue sunglasses standing on the grass in front of the Sydney Opera House holding a sign on the chest that says Welcome Friends!



A green sign that says "Very Deep Learning" and is at the edge of the Grand Canyon. Puffy white clouds are in the sky.



BIRD-SQL Benchmark

AT&T #1

Leaderboard - Execution Accuracy (EX)

| | Model | Code | Size | Oracle Knowledge | Dev (%) | Test (%) |
|-------------------|--|------------------------|------|------------------|---------|--------------|
| | Human Performance <i>Data Engineers + DB Students</i> | | | ✓ | | 92.96 |
| 1 Nov 11, 2024 | DSAIR + GPT-4o <i>AT&T - CDO</i> | | UNK | ✓ | 74.32 | 74.12 |
| 2 Nov 3, 2024 | CHASE-SQL + Gemini <i>Google Cloud</i> [Pourreza et al. '24] | | UNK | ✓ | 73.14 | 74.06 |
| 3 Oct 27, 2024 | ExSL + granite-34b-code <i>IBM Research AI</i> | | 34B | ✓ | 72.43 | 73.17 |
| 4 Aug 21, 2024 | OpenSearch-SQL, v2 + GPT-4o <i>Alibaba Cloud</i> | | UNK | ✓ | 69.30 | 72.28 |
| 5 Jul 22, 2024 | Distillery + GPT-4o <i>Distyl AI Research</i> [Maamari et al. '24] | | UNK | ✓ | 67.21 | 71.83 |
| 6 May 21, 2024 | CHESS _{IR+CG+UT} <i>Stanford</i> [Talaie et al.'24] | [link] | UNK | ✓ | 68.31 | 71.10 |
| 7 Aug 28, 2024 | Insights AI <i>Uber Freight</i> | | UNK | ✓ | 72.16 | 70.26 |
| 8 Aug 30, 2024 | PURPLE + RED + GPT-4o <i>Fudan University + Transwarp Technology</i> | | UNK | ✓ | 68.12 | 70.21 |
| 9 Jul 14, 2024 | RECAP + Gemini <i>Google Cloud</i> | | UNK | ✓ | 66.95 | 69.03 |
| 10 Jul 2, 2024 | ByteBrain <i>ByteDance Infra Lab</i> | | 33B | ✓ | 65.45 | 68.87 |

Overview

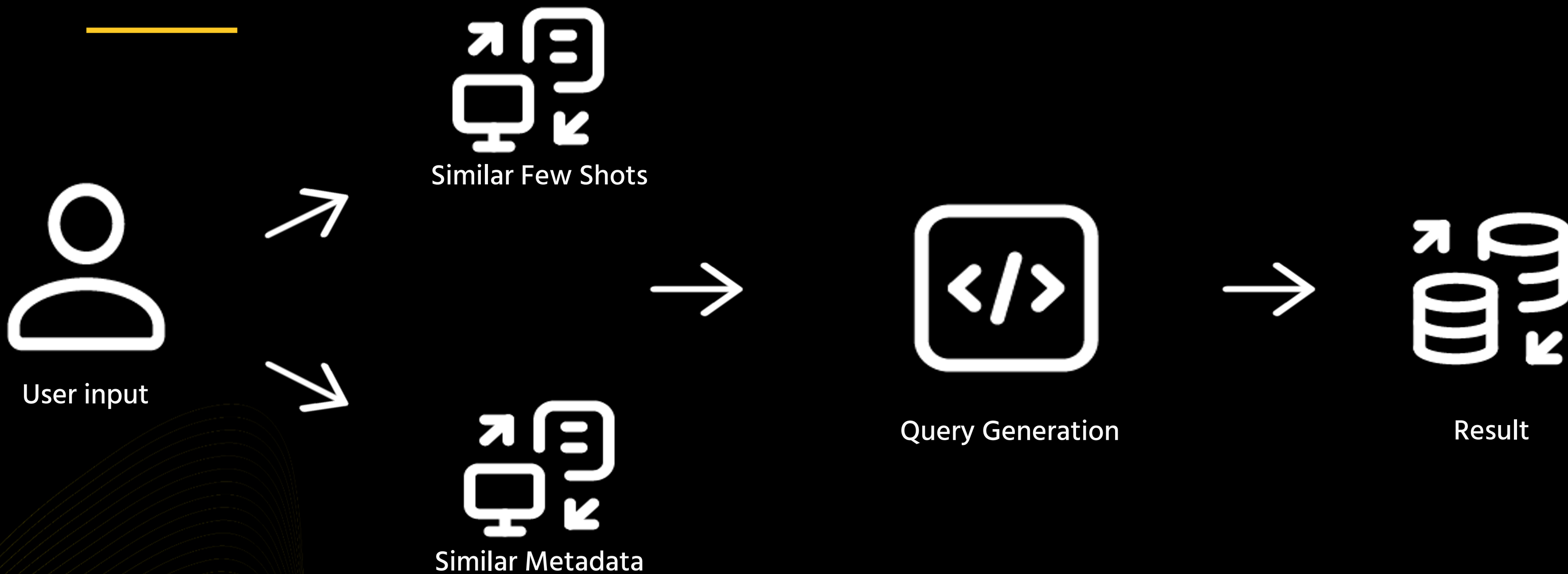


Number of prospect customers that are fiber eligible, limit to residential, exclude employees, exclude vacant addresses

23,410



Methodology



Methodology

Similar metadata



| Table Name | Column Name | Column Type | Column Description |
|------------------|----------------|-------------|--|
| acct_denorm | acct_id | numeric | Unique Identifier of an account |
| acct_denorm | wrls_pstpd | text | Indicates postpaid customers |
| customer_addr_ro | att_fiber_cvrg | text | Indicates if the location has fiber coverage |
| customer_addr_ro | prspct_cvrg | text | Indicates if the location is a prospect |
| customer_wirls | ipad_ind | text | Identifies a line/subscriber that has iPad |
| customer_wirls | dep_am | numeric | Identifies the initial deposit amount paid towards a device |
| acct_deno | spnh_bl_ind | text | Identifies an account whose owner has asked to have their bill sent to them in Spanish |
| customer_wirls | equip_nm | text | Identifies current equipment/device model name |
| customer_ro | dvc_wearable | text | Identifies if the account has an wearable device |
| customer_wirls | suspends_cnt | numeric | Number of times a subscriber has been suspended in past 12 months |



Methodology

Similar few shots



Number of prospect customers that are fiber eligible, limit to fiber new green in past 30 days, exclude employees

```
SELECT COUNT (DISTINCT r.addr_id)
FROM customer_address_ro r
JOIN customer_address a
    ON a.addr_id = r.addr_id
WHERE r.prspct_cvrg = 'Y'
    AND r.fiber_cvrg = 'Y'
    AND a.fiber_green_dt >= Current_date - 30
    AND r.att_empl = 'N';
```



Methodology



Query Generation

Generate a SQL query to answer {**user_question**}

Instructions
{**instructions**}

SQL examples
Provided are some examples with question and corresponding SQL codes
{**few_shots**}

Database Schema
The query will run on a database with the following schema:
{**metadata**}

Answer
Given the database schema, here is the SQL query that answers {**user_question**}



Fine Tuning



Fine Tuning Data Preparation

Data Curation

Ambiguous or Abbreviated Names:

Number of **con** that are **byod**

Number of prospect **aloc** that **aia** eligible

Number of **iru** or **cru** postpaid subscribers

Number of active **dsl customers** who are **upgrade** eligible

Number of customers who have upgraded their **hsia**



Fine Tuning Data Preparation

Data Curation

Consistency:

```
SELECT count(distinct srv_accs_id) FROM customer_wirls WHERE upgrd_eligible = 'NoTTCE'
```

```
select count(distinct srv_accs_id) FROM customer_wirls Where upgrd_eligible = 'OffCo'
```

```
Select count(distinct srv_accs_id) FROM customer_wirls where upgrd_eligible = 'PreCRO'
```



Fine Tuning Data Preparation

Data Curation

| Before | After |
|----------|----------|
| select | SELECT |
| count | COUNT |
| having | HAVING |
| where | WHERE |
| and | AND |
| or | OR |
| join | JOIN |
| limit | LIMIT |
| from | FROM |
| on | ON |
| in | IN |
| between | BETWEEN |
| limit | LIMIT |
| as | AS |
| distinct | DISTINCT |
| in | IN |



Fine Tuning Data Preparation

Data Curation

Unnecessary Parts:

```
SELECT COUNT(DISTINCT cust_loc_id) AS eligible_customers FROM ACCT_DENORM
```

```
SELECT COUNT(DISTINCT acct_id) FROM CUSTOMER_SUBSRPTN_WIRLS LIMIT 100
```

Table vs View Name:

```
SELECT COUNT(DISTINCT acct_id) FROM VIEWS_ACCT_DENORM
```

```
SELECT COUNT(DISTINCT account_id) FROM ACOUNT_DENORM
```



Fine Tuning Data Preparation

Data Profiling

actvtn_clstr VARCHAR(50), -- Activation Cluster name. The column values are a state name followed by Cluster like **California** Cluster.

equip_mdlnm VARCHAR(50), -- device model name associated with the line. The column values are **LG** for LG devices, **SM-%** or **%Galaxy%** or **Samsung** for Samsung devices, and **iPhone XX%** or **iPhone X%** for iPhone devices



Fine Tuning Data Preparation

Synthetic Data Generation



Introduction

You are given question, metadata, and query from MGO domain which is about telecommunication.

Goal

Your task is to generate synthetic data for fine tuning a large language model (LLM)

Instruction

{instruction}

Example for input and output

{examples}



Llama3-SQLcoder-8b Model Card

```
CREATE TABLE sales (  
  sale_id INTEGER PRIMARY KEY, -- Unique ID for each sale  
  product_id INTEGER, -- ID of product sold  
  customer_id INTEGER, -- ID of customer who made purchase  
  salesperson_id INTEGER, -- ID of salesperson who made the sale  
  sale_date DATE, -- Date the sale occurred  
  quantity INTEGER -- Quantity of product sold  
);  
  
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  
);  
  
-- 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
```






H2O LLM Studio



Dashboard

 H2O LLM Studio
v1.12.1

Navigation

- Home
- Settings

Datasets


- Import dataset
- View datasets

Experiments

- Create experiment
- Create grid search
- View experiments

Experiments

15



| Status | Count |
|------------------|-------|
| finished | 15 |
| queued + running | 0 |
| failed + stopped | 0 |

0.0%
Current GPU load

0.7%
Current CPU load

3.5 GB / 216.3 GB
Memory usage

Disk usage

44.20 %
222.1 GB / 502.9 GB

Detailed GPU stats

GPU #1 - current utilization: 0.0% - VRAM usage: 0.0 GB / 80.0 GB - NVIDIA A100 80GB PCIe

Grid Search

i Learning Rate is a grid search hyperparameter.

Learning Rate (grid search)

0.0001

Differential Learning Rate Layers

Select optional layers...

Attention Implementation

Auto

i Batch Size is a grid search hyperparameter.

Batch Size (grid search)

2

Select All | Deselect All

i Epochs is a grid search hyperparameter.

Epochs (grid search)

1

Select All | Deselect All

Schedule

Cosine

Min Learning Rate Ratio



0

i Warmup Epochs is a grid search hyperparameter.

Warmup Epochs (grid search)

0

i Weight Decay is a grid search hyperparameter.

Weight Decay (grid search)

0

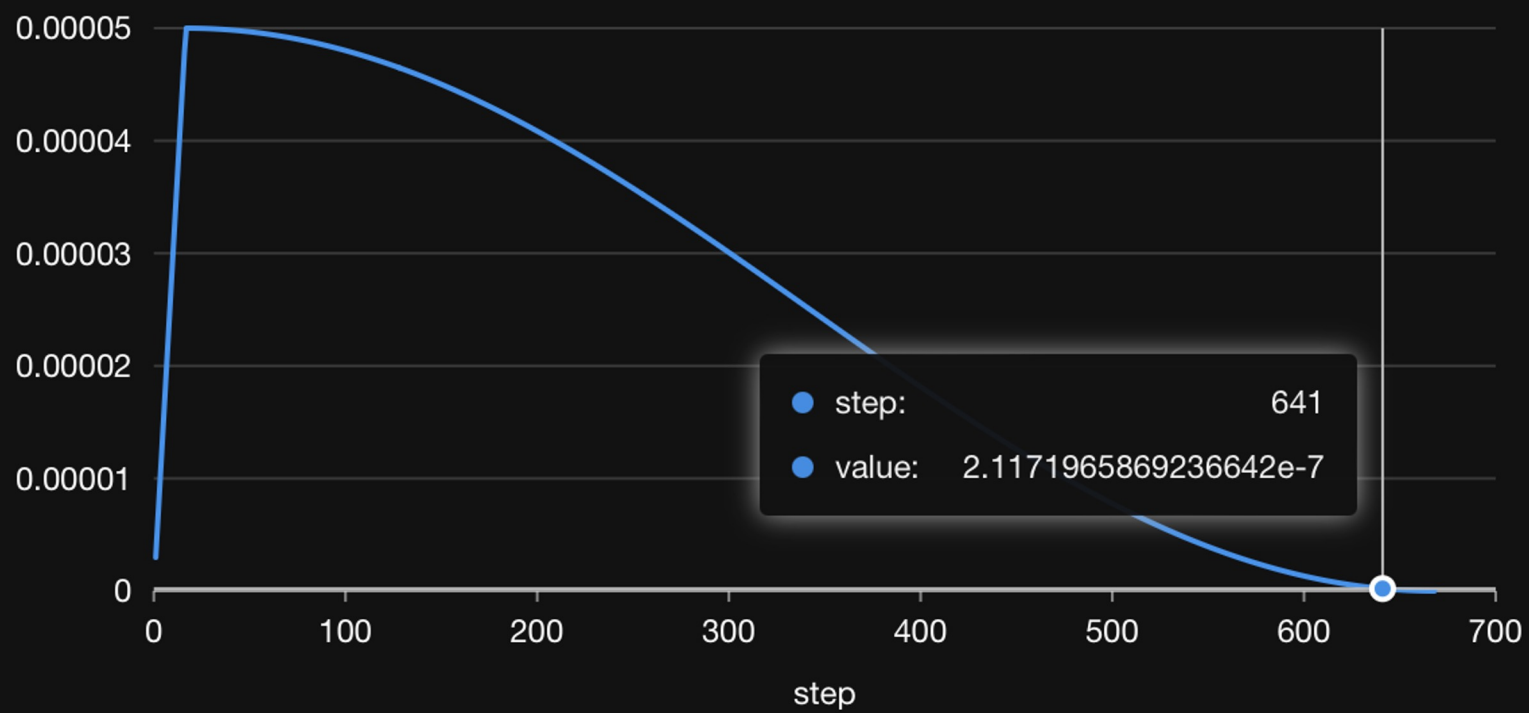
Logs

```
2024-11-11 13:01:53,639 - INFO: train loss: 0.45: 5%|5 | 8/160 [00:21<06:54, 2.73s/it]
2024-11-11 13:02:05,668 - INFO: train loss: 0.45: 5%|5 | 8/160 [00:33<06:54, 2.73s/it]
2024-11-11 13:02:14,431 - INFO: train loss: 0.32: 10%|# | 16/160 [00:42<06:21, 2.65s/it]
2024-11-11 13:02:25,669 - INFO: train loss: 0.32: 10%|# | 16/160 [00:53<06:21, 2.65s/it]
2024-11-11 13:02:36,315 - INFO: train loss: 0.13: 15%|#5 | 24/160 [01:04<06:05, 2.69s/it]
2024-11-11 13:02:51,817 - INFO: train loss: 0.13: 15%|#5 | 24/160 [01:20<06:05, 2.69s/it]
2024-11-11 13:02:57,252 - INFO: train loss: 0.17: 20%|## | 32/160 [01:25<05:40, 2.66s/it]
2024-11-11 13:03:11,818 - INFO: train loss: 0.17: 20%|## | 32/160 [01:40<05:40, 2.66s/it]
2024-11-11 13:03:18,438 - INFO: train loss: 0.11: 25%|##5 | 40/160 [01:46<05:18, 2.66s/it]
2024-11-11 13:03:31,819 - INFO: train loss: 0.11: 25%|##5 | 40/160 [02:00<05:18, 2.66s/it]
2024-11-11 13:03:39,391 - INFO: train loss: 0.15: 30%|### | 48/160 [02:07<04:56, 2.64s/it]
2024-11-11 13:03:51,820 - INFO: train loss: 0.15: 30%|### | 48/160 [02:20<04:56, 2.64s/it]
2024-11-11 13:04:00,334 - INFO: train loss: 0.06: 35%|###5 | 56/160 [02:28<04:34, 2.64s/it]
2024-11-11 13:04:11,820 - INFO: train loss: 0.06: 35%|###5 | 56/160 [02:40<04:34, 2.64s/it]
2024-11-11 13:04:21,330 - INFO: train loss: 0.12: 40%|#### | 64/160 [02:49<04:12, 2.63s/it]
2024-11-11 13:04:31,821 - INFO: train loss: 0.12: 40%|#### | 64/160 [03:00<04:12, 2.63s/it]
2024-11-11 13:04:43,036 - INFO: train loss: 0.12: 45%|####5 | 72/160 [03:11<03:53, 2.66s/it]
2024-11-11 13:04:55,672 - INFO: train loss: 0.12: 45%|####5 | 72/160 [03:23<03:53, 2.66s/it]
2024-11-11 13:05:03,866 - INFO: train loss: 0.09: 50%|##### | 80/160 [03:32<03:31, 2.64s/it]
2024-11-11 13:05:15,673 - INFO: train loss: 0.09: 50%|##### | 80/160 [03:43<03:31, 2.64s/it]
2024-11-11 13:05:24,882 - INFO: train loss: 0.12: 55%|#####5 | 88/160 [03:53<03:09, 2.64s/it]
2024-11-11 13:05:35,674 - INFO: train loss: 0.12: 55%|#####5 | 88/160 [04:03<03:09, 2.64s/it]
2024-11-11 13:05:46,269 - INFO: train loss: 0.04: 60%|##### | 96/160 [04:14<02:49, 2.65s/it]
2024-11-11 13:06:01,823 - INFO: train loss: 0.04: 60%|##### | 96/160 [04:30<02:49, 2.65s/it]
2024-11-11 13:06:07,915 - INFO: train loss: 0.07: 65%|#####5 | 104/160 [04:36<02:29, 2.67s/it]
2024-11-11 13:06:21,824 - INFO: train loss: 0.07: 65%|#####5 | 104/160 [04:50<02:29, 2.67s/it]
2024-11-11 13:06:29,597 - INFO: train loss: 0.14: 70%|##### | 112/160 [04:57<02:08, 2.68s/it]
2024-11-11 13:06:41,825 - INFO: train loss: 0.14: 70%|##### | 112/160 [05:10<02:08, 2.68s/it]
2024-11-11 13:06:51,620 - INFO: train loss: 0.06: 75%|#####5 | 120/160 [05:19<01:48, 2.70s/it]
```

Charts

Charts Summary Train Data Insights Validation Prediction Insights Logs Config Chat

LEARNING RATE ⓘ



TRAIN BATCH LOSS ⓘ

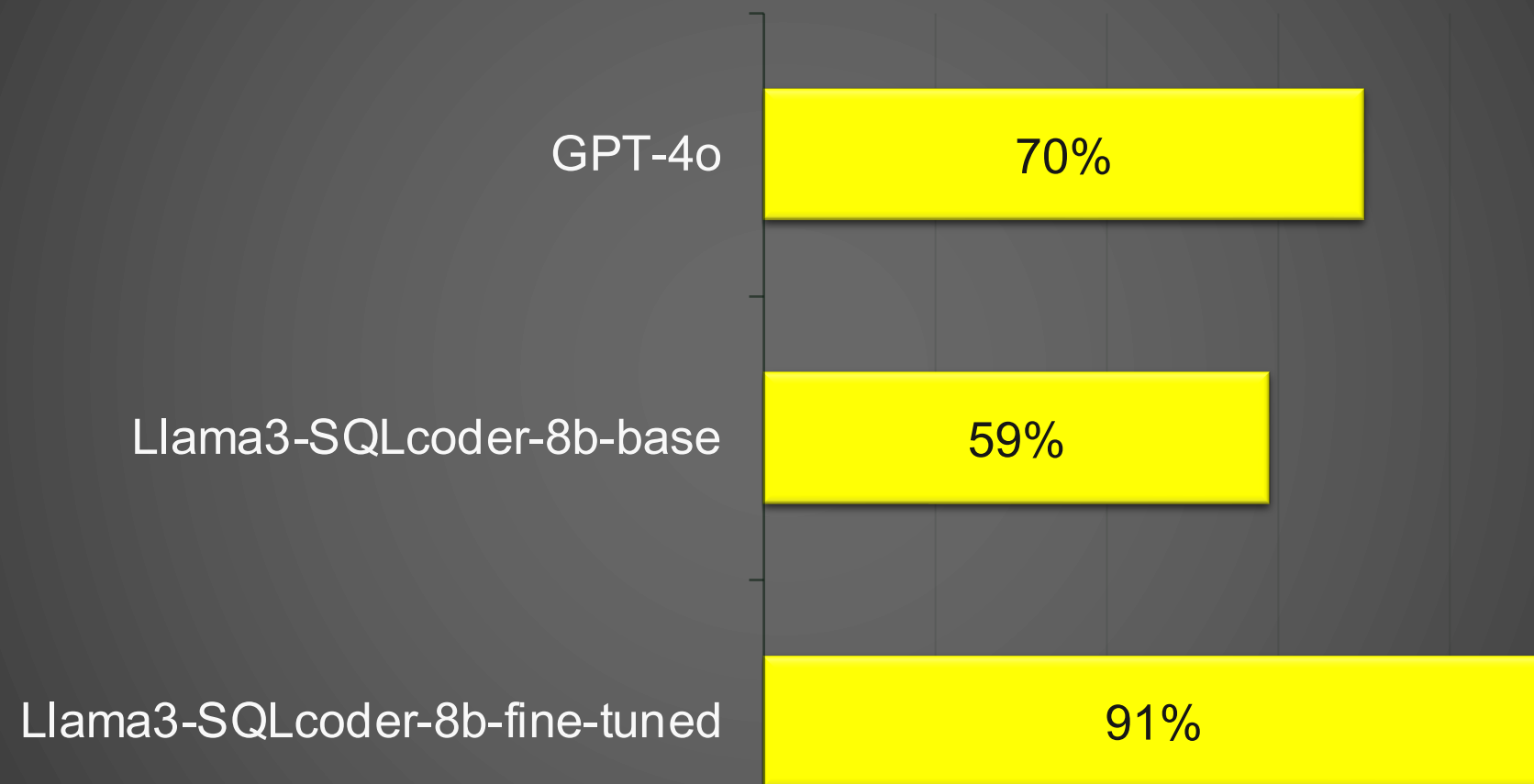


Results

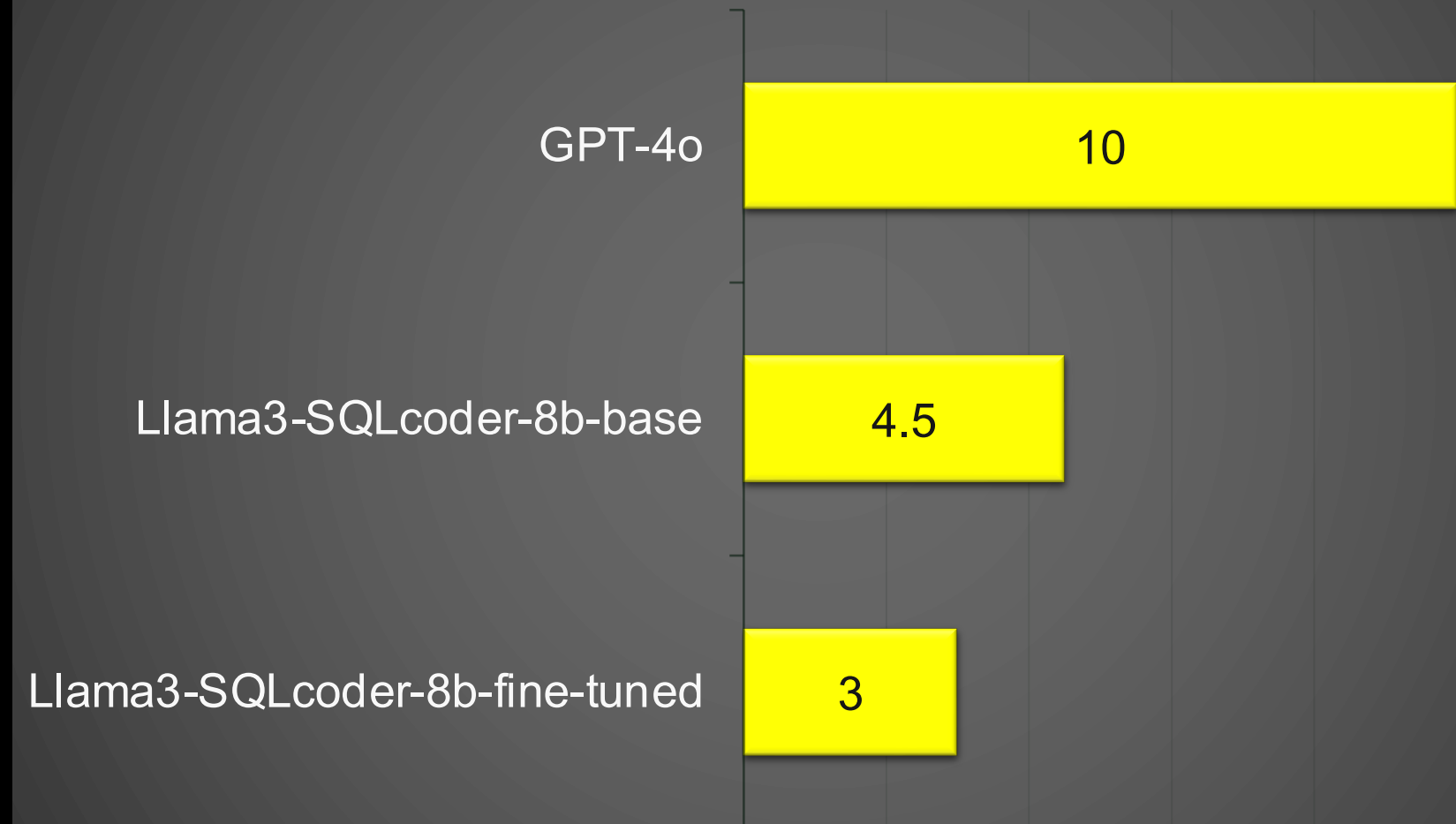


Results

Query Generation Accuracy (%)



Query Generation Latency (Seconds)



GPT-4o costs around **\$0.02** per call while Llama3 is effectively **free**





MAKERS

Thank you!

