World's leading telco leverages H2O LLM Studio to finetune models and enhance query generation with reduced costs

AT&T

H2O.ai

OVERVIEW:

In today's competitive landscape, collecting actionable insights from databases is critical for business units looking to enhance their campaigns and drive growth. Recognizing this need, the CDO Office team at AT&T aimed to create a user-friendly tool that empowered non-technical users to interact with their database using plain English and getting results back in numbers, enabling users to uncover valuable insights, such as identifying key individuals in specific regions to target for campaigns and market expansion. For example, a business user might want to know, "Number of prospective customers that are fiber-eligible, limited to residential, excluding employees and vacant addresses," and instantly access critical data. By simplifying complex database interactions, the tool democratizes data-driven decision-making across the organization

CHALLENGES:

The business unit needed a solution that not only identified prospective customers but also analyzed current customers to uncover untapped opportunities—services the company provides but the customers hadn't purchased yet. Precision was critical, as the insights derived from the solution would drive key decisions, leaving no room for error.

While a large language model provided the required analytical capabilities, its high operational costs posed a challenge. To address this, the team explored how to fine-tune a smaller language model, aiming to achieve comparable performance at a significantly lower cost, ensuring both accuracy and efficiency, using H2O LLM Studio.

WHY FINE-TUNING?

The team opted for fine-tuning a small language model instead of using a Retrieval-Augmented Generation (RAG) approach that relies on a large language model for two key reasons: cost reduction and improved latency. Fine-tuning allowed them to maintain high performance while achieving faster response times at a fraction of the cost.

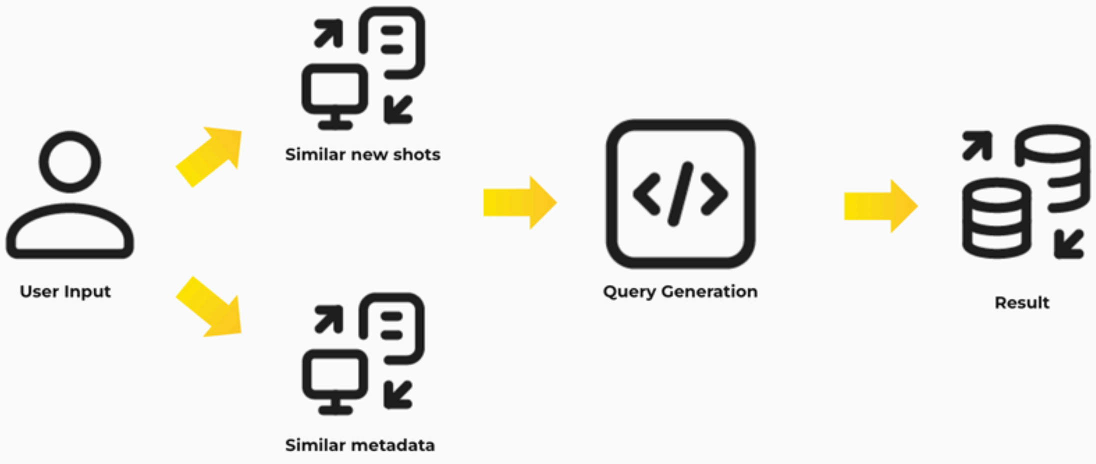
As Farbod Tavakkoli, data scientist at AT&T, explains, “While a small language model may not necessarily outperform a large language model, some tasks are simple enough that a small model can achieve the same quality or accuracy with significantly lower costs.”

He added that for those interested in benchmarking text-to-SQL solutions, tools like the BIRD-SQL benchmark, widely used by organizations and institutions such as Google, IBM, Alibaba, and Stanford, offer a way to compare enterprise solutions against global standards, evaluate and validate their text-to-SQL capabilities. AT&T secured #1 spot in this global text-to-SQL benchmark and is currently ranked #2.

METHODOLOGY

The team enhances LLM query generation through **three steps:**

- ✓ **User Input and Semantic Search—questions are matched with similar queries and relevant metadata from the database schema;**
- ✓ **LLM Query Generation—this information guides the model in creating a single query;**
- ✓ **Database Search and Execution—the query identifies the most relevant table and column data, retrieves results, and executes efficiently. While this approach excels with larger models like GPT-4, fine-tuning smaller models balances quality with cost savings.**



FINE-TUNNING STEP-BY-STEP:

- To fine-tune the model, the team starts with data curation, which involves preparing the data for the model. This step is crucial in teaching the model the terminology and keywords used on a daily basis.

- After data curation, the team provides extra information about columns within the database, allowing the model to understand the context. It's essential to avoid synthetic data generation due to legal barriers and instead invest in obtaining the necessary data from the business unit.

- The team selects a suitable model for fine-tuning, in this case, the Llama-Transformer SQL coder 8 billion parameter model. The model is fine-tuned using the curated data and data profiling information.

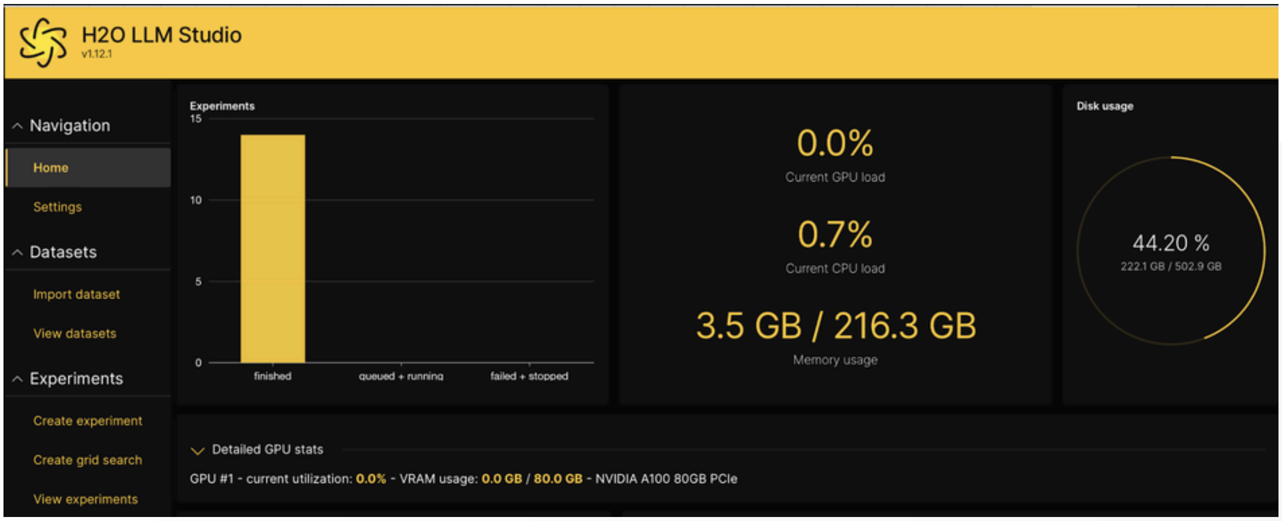
- Using H2O LLM Studio, the team conducts a grid search to find the optimal parameters for the model. This step is crucial in identifying the right learning rate and number of epochs for the model

- The team tracks the model's performance during fine-tuning, monitoring the training logs to ensure convergence. If the logs do not converge, the team can stop the process and investigate the issue.

- Finally, the team deploys the fine-tuned model and compares its performance to the initial solution using GPT4. The results show that the fine-tuned model is not only more cost-effective but also provides better performance.

WHY SMALL LANGUAGE MODELS AND H2O LLM STUDIO?

The methodology outlined demonstrates the effectiveness of fine-tuning small language models using LLM Studio as a cost-effective solution for natural language processing tasks. By leveraging the power of semantic similarity search, data curation, and data profiling, the AT&T CDO team was able to fine-tune a small language model to achieve comparable performance to a large language model like GPT 4.0.



0.0%
Current GPU load

0.7%
Current CPU load

3.5 GB / 216.3 GB
Memory usage

44.20 %
222.1 GB / 502.9 GB

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

RESULTS

The results show that fine-tuning a small language model with **H2O LLM Studio can significantly reduce costs**, with the Llama SQL coder 8 billion parameter model being effectively free to use, compared to the two cents per call cost of the previous solution.

Furthermore, fine-tuning small language models with LLM Studio provides a more tailored solution that **can be customized to meet the specific needs of a business or organization.** By using data curation and data profiling, the model can be trained to understand the terminology and keywords used within a specific domain, leading to more accurate and relevant results.

Overall, fine-tuning small language models with LLM Studio offers a cost-effective solution that provides a **high degree of customization and accuracy.**

HIGHLIGHTS:

Significant **Cost Reduction** › **Higher ROI**

Higher **Costumization** › **Optimized performance**

Maximized **Accuracy and efficiency** › **Faster decision making**