# Transforming Call Center Operations and Enhancing Customer Service in Telco with Small Language Models

## OVERVIEW:

**AT&T, a renowned broadband connectivity provider and H2O.ai customer, receives 5 million customer calls annually, resulting in a vast amount of recorded, transcribed, and summarized interactions. To unlock the value of these conversations, AT&T is leveraging the power of AI and language models to drive innovation and improve customer service.**

**The company is exploring the intersection of AI and telecommunications by distilling large language models, such as GPT-4, into smaller, more economical models, such as H2O Danube. This approach enables AT&T to extract valuable insights from customer interactions, which are then used to enhance customer service and support.**

**Their key objectives include:**

- Early Conversation Insights: Identify patterns and themes in customer interactions to anticipate potential needs and proactively address issues before they escalate.
- Deeper Value Extraction: Uncover rich, detailed insights from conversation summaries to inform customer service and support strategies.
- Insight Extraction: In order to extract insights from customer interactions, the company uses multi-label classification for 80 categories related to customer service and support issues. Examples include: Product pain points, agent actions and effectiveness, customer sentiment, complaints about poor performance, rescheduling appointments, and assistance with moving.

## CHALLENGES:

The initial solution with GPT-4 produced qualified outputs and they were able to save 50,000 customers annually. However, they were facing a few challenges:

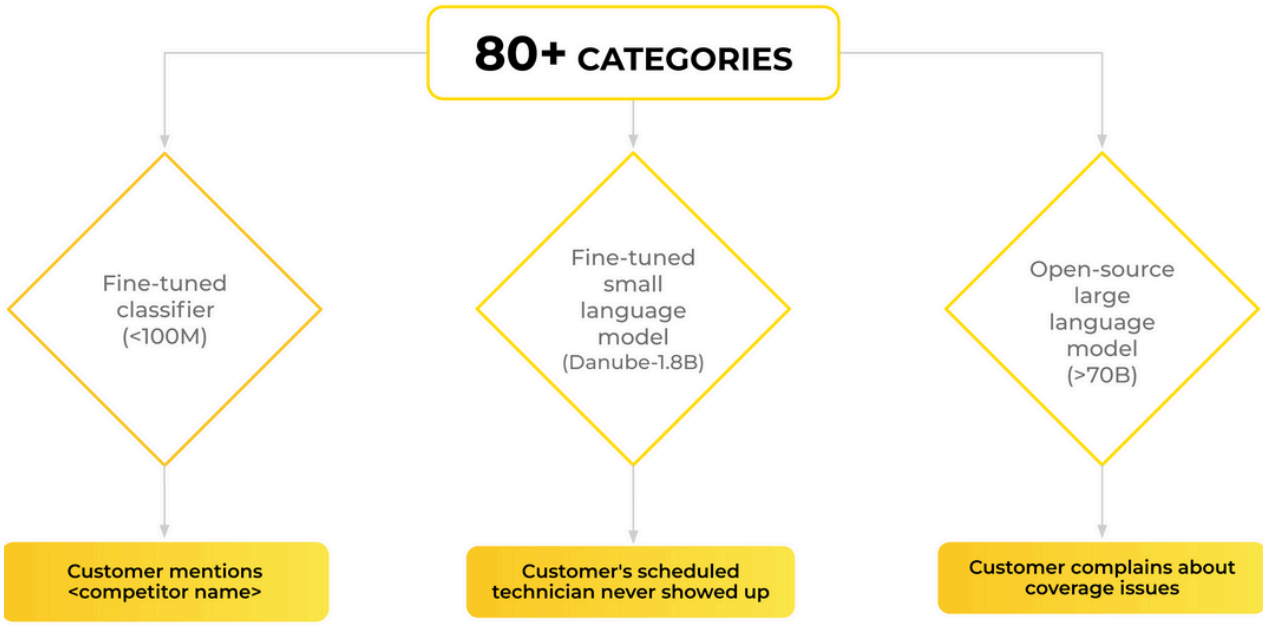
- The cost was getting too high.
- It took approximately 15 hours to process one day's worth of categorization so it wasn't possible to scale.
- Because the vendor from the initial solution wasn't included in their subscription, they were encountering privacy and security issues.

# SOLUTION POWERED BY H2O.AI'S SLM:

AT&T decided to distill large language models into three smaller fine-tuned open source models to reduce computational costs and improve scalability. This approach enables AT&T to deploy AI-powered language models in a production environment, driving innovation and business value.

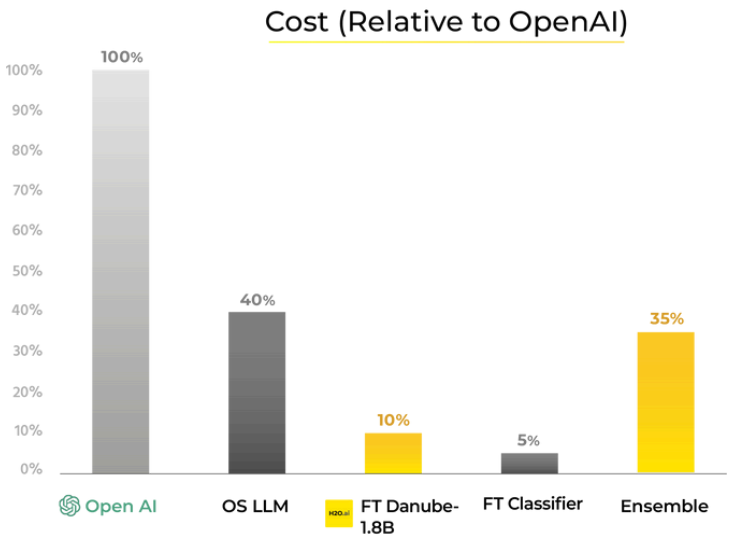
The first model, a fine-tuned classifier, was able to cover 50 from the 80 categories within their multi-label classification. The next solution for the remaining cases was our fine-tuned small language model named Danube 1.8B. H2O LLM Studio was used to fine-tune this model and it was able to capture 20 of the categories. Example: "customers scheduled and technicians never showed up".

For the remaining 10 categories, the team opted to use an open-source large language model. For this one, they used Llama 3.170B. Example: "customer complaints about coverage issues".



## BUSINESS VALUE:

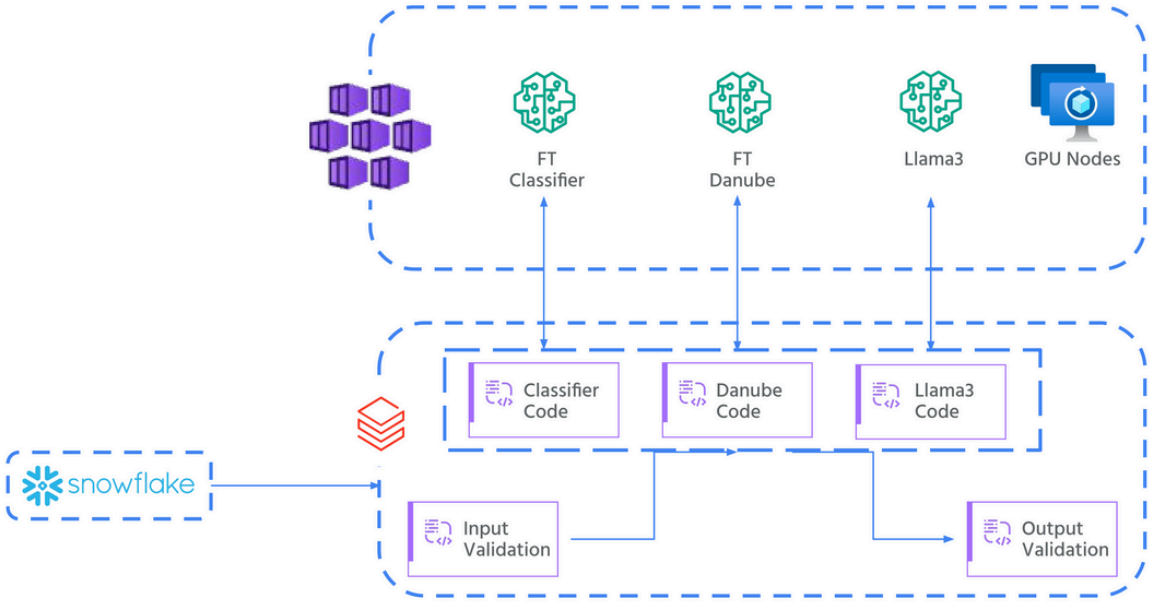
The team working on the project was able to get 91% accuracy from the combined performance of new models, which is very close to the previous and more expensive solution. Also, by taking the GPT-4 cost as 100%, their open source model did about 42%. Danube was 10% in their fine-tuned classifier. But because they took 10 questions from Llama, 20 questions from Danube, and 50 questions from the classifier, they achieved 35% cost relative to Open AI's.




## FINE-TUNING MODEL IMPLEMENTATION:

The company utilizes Snowflake's database as its data source and a Databricks workspace consisting of five notebooks. The first notebook aggregates, filters, and prepares the data for processing. The prepared data is then passed to three separate notebooks, each dedicated to a specific model. However, instead of sending the entire dataset to each model, the model itself filters the data and only sends the required categories to the AKS GPU node cluster for processing.

This targeted approach is crucial, as it avoids sending all 80 categories to the model. Once the models, which are running on the AKS GPU clusters, have completed their tasks, the outputs are combined in the final notebook. This last notebook performs filtering, validation, and saves the processed data to the target data point.



## RESULTS:

Compared to their previous solution, the new approach yielded significant improvements. The most notable difference was the substantial **reduction in processing time, from 15 hours to just 4.5 hours**.

Additionally, the new workflow utilizes three models, a significant upgrade from the single model used previously. This design also allows for seamless integration of new models, enabling easy adoption of future advancements in accuracy and performance.

Notably, AT&T achieved a substantial increase in **transcript processing capacity, growing from 45,000 to 250,000.**

Overall, the new solution delivered a triple benefit: time savings, increased transcript processing, and reduced costs.

## HIGHLIGHTS:

**75%** latency improvement → **Faster time to value**

**500%** scalability → **Business Velocity**

**70%** cost reduction → **Higher ROI**

## UNLOCKING THE POWER OF SMALL LANGUAGE MODELS WITH H2O.AI:

Like most decisions in AI and tech, the decision of which Language Model to use for your production use cases comes down to trade-offs. Many Large Language Models are excellent at solving a wide-range of Natural Language Understanding use cases out-of-the-box but require sending your data to a third party who is hosting the model, or having enough GPUs to run the large model privately. On the other hand, Small Language Models are more compact, nimble, and take significantly fewer resources to deploy - these smaller models are great at handling specific tasks and can be tailored to perform well within a narrower scope. When you have an application where resources are scarce and your use case is well defined, SLMs can be the way to go.

H2O DANUBE

The H2O Danube series consists of compact foundational models, each containing 1.8 billion parameters. These models are designed to be lightweight and quick to respond, making them suitable for a broad array of use cases, including: retrieval augmented generation, open-ended text generation, brainstorming, summarization, table creation, data formatting, paraphrasing, extraction, chat, and more.

H2O LLM STUDIO

H2O LLM Studio, in turn, comes into play when users need a custom solution. Created by our top Kaggle Grandmasters, this no-code fine-tuning framework empowers organizations to build their own state-of-the-art Large Language Models for enterprise applications.