LTIMindtree Achieves a Double-Digit Performance Improvement Using SigOpt for AI Model Tuning

LTIMindtree used the SigOpt Intelligent Experimentation Platform to operationalize the production of AI models for better end customer results.

Using SigOpt yielded up to 63 percent improvement in inference time beyond baseline performance for pretrained AI transformers.

As of November 14, 2023, Mindtree merged with L&T Infotech to form LTIMindtree. LTIMindtree is a global technology consulting and digital solutions company with decades of deep engineering experience, specializing in the delivery of agile and comprehensive enterprise solutions that use novel artificial intelligence (AI) technology. One of the cornerstones in the LTIMindtree AI suite of products is its AI-powered call summarizer, which helps call center agents summarize completed customer calls, a complex task that requires AI models to be both current and consistently improved at a granular level. LTIMindtree is committed to continuous improvement of AI model performance and response times to ensure superior products for their customers. In this case, the goal was to enable faster and more accurate call summaries using AI tools.

For LTIMindtree, improving call summarization meant meticulously fine-tuning AI models for use with individual customers. However, manually fine-tuning AI models for *every* individual customer was time consuming and introduced operational bottlenecks to the process. The LTIMindtree team sought a tool to overcome these challenges and settled on using SigOpt to automate the entire fine-tuning workflow.

intel. + SIGOPT + LTIMindtree

> *"The investment with the SigOpt Intelligent Experimentation Platform was definitely worth the results LTIMindtree has achieved."*
>
> Bhanu Prakash Aladahallinanjappa, program architect, LTIMindtree[1]

By using SigOpt, a world-class solution for operationalizing the production of AI models, LTIMindtree saw not just a reduction in the operational cost but also a significant performance gain in both accuracy and inference time. This SigOpt-enabled workflow ultimately translated to an up to 63 percent reduction in inference time when compared with the performance of the pre-trained HuggingFace transformer BART.[2] More importantly, this model performance gain translated to a better, more productive customer experience.

LTIMindtree is focused on delivering operational efficiency by using avant-garde AI models tailored to customer needs and data. In this case, LTIMindtree's expertise was intended for application in the telecommunications industry, where AI-enabled chatbots can help reduce voluntary customer churn, increase average revenue per user, and deliver cost-effective customer convenience. When customers interact with a chatbot, it needs to be a positive, efficient, and accurate experience for long-term customer retention and minimal need for human assistance.[3]

Chatbots help provide better customer experiences through quick resolutions while increasing employee efficiency and productivity. Speed, responsiveness, and accuracy are key to improving customer service. In a telecom, as well as in most other customer service environments, call accuracy is measured by the Recall-Oriented Understudy for Gisting Evaluation (ROUGE) score.

Comparing SigOpt to other methods

Conventional hyperparameter tuning methods like grid search and random search are tedious and time consuming. LTIMindtree was confronted with this challenge when fine-tuning hyperparameters to improve chatbot performance in call responses and summaries. Specifically, they wanted to select the right deep learning architecture to give chatbot exchanges a measurable improvement in performance. Through the following experiments, LTIMindtree sought to determine whether the SigOpt Intelligent Experimentation Platform was the right tool to facilitate the fine-tuning of hyperparameters.

Rather than manually interpreting and optimizing its AI models or combining multiple tools within LTIMindtree's existing infrastructure, the SigOpt Intelligent Experimentation Platform gave LTIMindtree step-by-step clarity with a model-agnostic cloud-based solution.

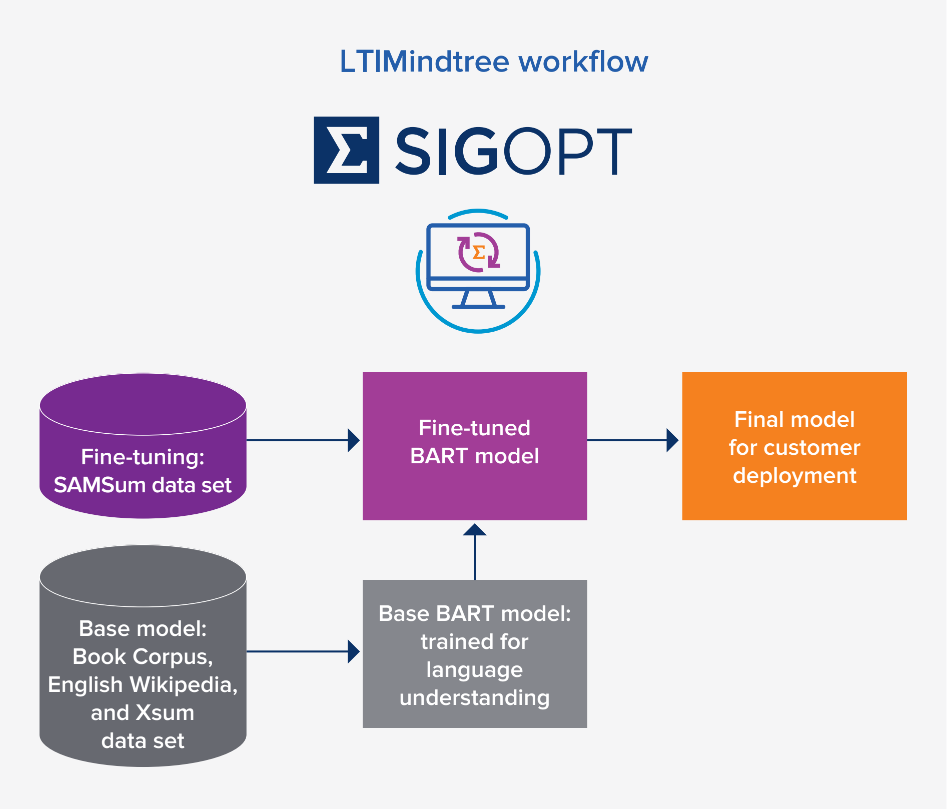
LTIMindtree workflow
SIGOPT
Fine-tuning: SAMSum data set
Fine-tuned BART model
Final model for customer deployment
Base model: Book Corpus, English Wikipedia, and Xsum data set
Base BART model: trained for language understanding

Figure 1. LTIMindtree's workflow for fine-tuning AI models using the SAMSum data set.

SigOpt optimizes AI models through a combination of Bayesian and other global optimization algorithms to boost model performance while reducing costs and saving time. With SigOpt, LTIMindtree experts implemented their model optimization experiments with a few lines of code using LTIMindtree's existing environments and tools. SigOpt brings together data management, optimization, analytics, transparency, and scalability so modelers can track runs, visualize training curves, and optimize hyperparameters via an integrated dashboard designed exclusively for intelligent experimentation. LTIMindtree could easily see what did and didn't work within their experiments, as well as a history of what had been done and how the different models were performing.

> *"Model customization and accuracy improvements are what our customers are insisting on instead of using available cloud solutions in the contact center space. We will definitely recommend or offer our service around customization using the SigOpt Intelligent Experimentation Platform."*
>
> Bhanu Prakash Aladahallinanjappa,
> program architect, LTIMindtree

Before tackling the project, LTIMindtree chose to become familiar with SigOpt by comparing the SigOpt optimization algorithms to other standard approaches for hyperparameter optimization.

For the comparison, LTIMindtree used grid search, random search, and the SciPy optimization library and compared performance on the eggholder function. The eggholder function is considered a classic function within the optimization literature. The resulting comparison gave quick insight into how SigOpt was able to find the best value in fewer iterations compared to the other approaches.

Method	Number of evaluations	Max value achieved
Bayesian optimization by SigOpt tool	40	158.55
Grid search	80	138.22
Random search	70	145.95

> *"LTIMindtree's cognitive contact center solution now offers unprecedented accuracy and performance of the summarization model to customers, courtesy of the features in the SigOpt Intelligent Experimentation Platform"*
>
> Lakshmi Ranganathan,
> technical consulting engineering lead, Intel India

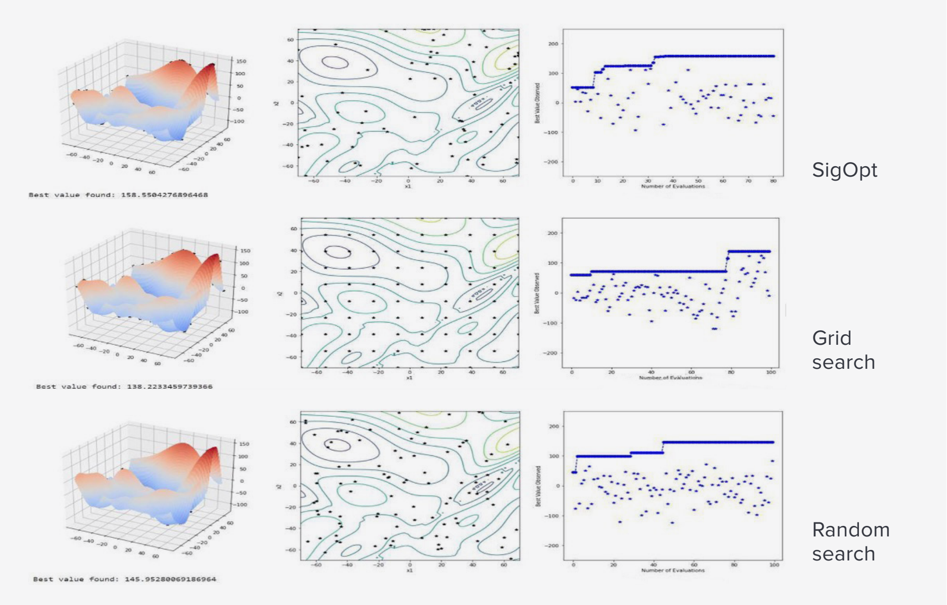

Figure 2. The LTIMindtree team tried conventional tuning methods like grid search and random search. This image displays the results comparison with SigOpt.[1]

For additional validation, LTIMindtree tried image classification on the MNIST data set using a multilayer perceptron model.

Overall, running these two test cases gave LTIMindtree the confidence to continue with SigOpt for this project.

Choosing the right deep learning architecture that aligns with modeling and business objectives

One of the challenges when building deep learning models is to choose the right architecture. Choosing the right architecture means finding the architecture that is not just ensuring the best model performance but is also satisfying the requirements for going into production. Once LTIMindtree verified the effectiveness of SigOpt, LTIMindtree used the platform to help build a case for determining which AI transformer model architecture to use for their customers.

To select the right transformer model architecture, LTIMindtree experimented using a GPU-enabled machine. To identify a model candidate, LTIMindtree tested a Pegasus model and a BART model, both from Hugging Face. They also chose the SAMSum Corpus data set for abstractive summarization.

LTIMindtree's goal was to fine-tune a pretrained model—optimizing batch size and learning rate—to understand if the SigOpt Intelligent Experimentation Platform could select new and better hyperparameter values and improve transformer models.

Initially, LTIMindtree tried to optimize the Pegasus model by minimizing evaluation loss. Seeing the early results gave the LTIMindtree team the confidence to rule out the Pegasus model as a potential model candidate.

An AI transformer model is a deep learning model commonly used in natural language processing tasks that adopts the mechanism of self-attention to weigh individual parts of the input data.

After ruling out the Pegasus model, LTIMindtree switched their attention to the BART model. They also switched from looking at evaluation loss to looking at the ROUGE score metrics to better compare the pretrained baseline models. The ROUGE score comes in three different forms:

- **ROUGE 1** – Number of matching unigrams in sentence/total unigrams in sentence
- **ROUGE 2** – Number of matching bigrams in sentence/total bigrams in sentence
- **ROUGE L** – Length of longest common subsequence/total words in sentence

The metrics measure the difference between the true sentence and the generated sentence—in this case, for summarization. As a baseline, LTIMindtree used the pretrained BART model and got a ROUGE 1 score of 54.39. This is also the model the team decided to use for weight initialization for the fine-tuning.

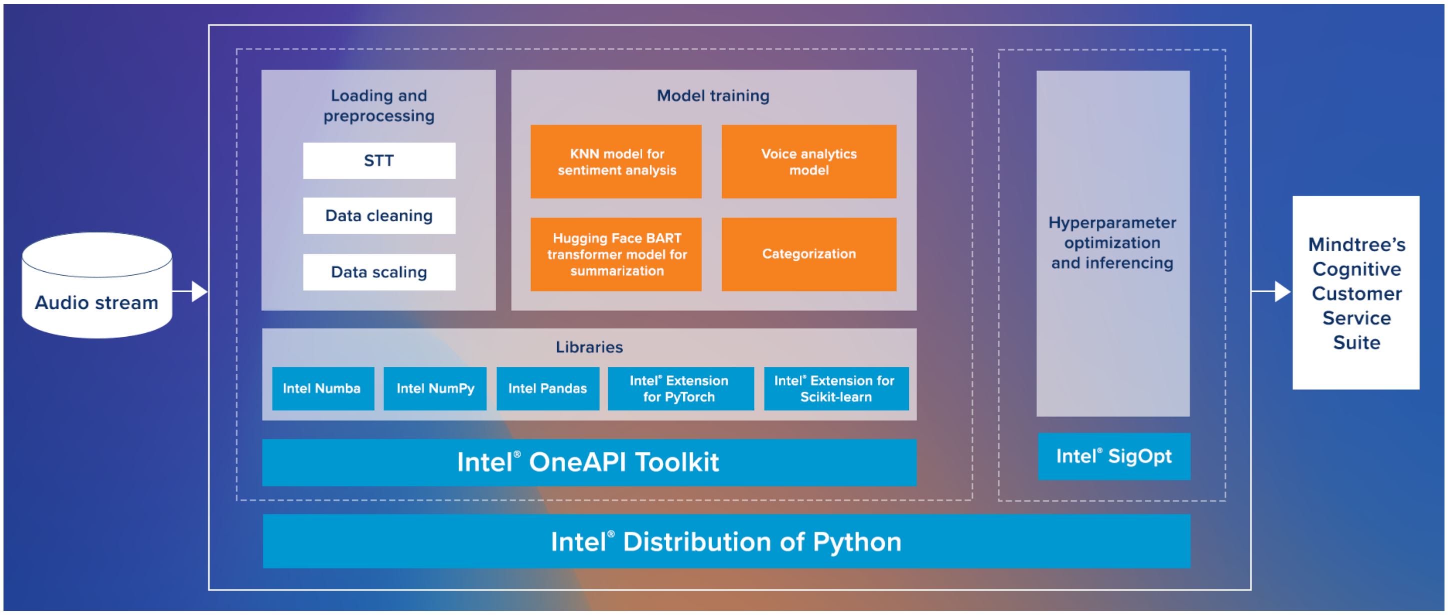
Audio stream
Loading and preprocessing
STT
Data cleaning
Data scaling
Model training
KNN model for sentiment analysis
Voice analytics model
Hugging Face BART transformer model for summarization
Categorization
Libraries
Intel Numba
Intel NumPy
Intel Pandas
Intel® Extension for PyTorch
Intel® Extension for Scikit-learn
Intel® OneAPI Toolkit
Hyperparameter optimization and inferencing
Intel® SigOpt
Intel® Distribution of Python
Mindtree's Cognitive Customer Service Suite

Figure 3. LTIMindtree's solution block diagram.

Operationalizing the production of AI models to optimize the end customer experience

For summarization of the call conversation, a pretrained Facebook/BART-large-XSum model was used as the base model. The model was fine-tuned on the SAMSum Corpus data set, a human-annotated dialogue data set for abstractive summarization. Hyperparameter optimization was carried out using SigOpt with the ROUGE score as the performance metric.

With dual goals of maximizing the ROUGE score and minimizing inference time, LTIMindtree used **SigOpt's multimeric optimization** advanced feature to visualize possible design solutions. Using SigOpt's web application, LTIMindTree was able to visualize how these two competing metrics performed after running multiple experiments and created a Pareto Frontier visualizing the best outcomes for each metric combination. From here, LTIMindtree was able to select the best configuration for their models to bring into production.

As a result of using SigOpt, LTIMindtree optimized its AI models to achieve 63 percent faster responsiveness and better call summarization through contact center bots. The models not only automate the task of postcall summarization but also provide text summaries that are consistent and accurate, thereby improving productivity and performance. Below is a summary from the baseline model and a summary from the model fine-tuned using SigOpt.

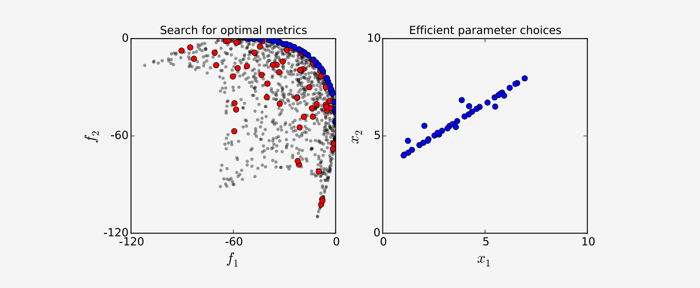

Figure 4. The Pareto Frontier provided by SigOpt's visualization tools

Input transcript conversation file

Agent: Thank you for Calling, ABC. My name is JERRY, how can I help you?

Customer: Yes, I have been using Apache Airflow. The scheduler is not working properly.

Agent: I'm sorry. Can I know what specific API is not working?

Customer: The API where you call the cron expression to schedule the pipeline to a specific cycle, is not working.

Agent: I'm really sorry for the inconvenience, let me transfer you to the API team, as they are the only one who can help you with the situation.

Customer: Ok.

API_team: Hi Sir, how can I help you?

Customer: You cron API is not working. This has resulted in a big loss for my company.

Agent: I'm sorry sir, can you please tell me that time zone you used and your country sir?

Customer: My company is in India, I used EST.

Agent: Sir as you are from India, you must use IST and not EST.

Customer: ok, wait let me check.

Agent: ok sir.

Customer: Yes, It is working now.

Agent: OK Sir, can I help you with anything alse?

Customer: NO no.. Thank you.

Agent: Thank you sir, have a nice day.

Summary generated by model *without* SigOpt suggestion

The scheduler in Apache Airflow is not working properly. This has resulted in a big loss for the customer's company. The team from API_team has fixed the problem. The customer is from India and his company is in EST time zone. The API that allows him to schedule the pipeline to a specific cycle is working now.

Summary generated by the model *with* SigOpt suggestion

The scheduler in Apache Airflow is not working properly.

The API that allows to schedule the pipeline to a specific cycle is broken.

Customer's company is in India and **he used EST time zone, so he needs to use IST**.

Figure 5. The summary from the fine-tuned model is more concise and efficient.[2]

> *"We intended to use the SigOpt Intelligent Experimentation Platform to improve model accuracy (ROUGE score), but we also got significant CPU performance benefits."*
>
> Bhanu Prakash Aladahallinanjappa, program architect, LTIMindtree

Improving model accuracy and realizing significant CPU performance benefits to boost inference by up to 63 percent

Getting a more concise reference report of customer engagements is a key success criterion for any abstractive summary model. On average, LTIMindtree found that manual summary extraction takes five to ten minutes, depending upon call length, and is subject to the potential for human error. Additionally, the increased speed through efficiency helps agents switch from one call to the next. Both the chatbot responsiveness and engagement summaries have now been vastly improved due to working with SigOpt.

SigOpt was able to find models that performed well and have lower inference times despite the industry trend for the opposite to occur. Keep in mind, also, the inference improvements shown here are multiplicative with hardware performance improvements. Meaning a more efficient processor will accelerate these gains even further. For reference, the Intel hardware configuration LTIMindtree used was the Intel® Xeon® Platinum 8380 Processor.[4] The following results show this reduced inference time.[5]

As a result of using SigOpt, LTIMindtree now has a framework that enables them to transfer knowledge from one project to the next.

Transcript files with summary	Inference time in seconds	Model name
Customer is angry because he can't access his email for almost a week. Rocket Speed Internet will call him back at the same number, but he needs to check which lights are lit on his modem. There is a problem with the DSL line behind customer's desk. It's the gray phone cord at the back part of the modem. The agent tells customer to plug it in and it fixes the problem.	16.75044	Hugging Face latest model; model created without SigOpt help
Customer can't access his email for almost a week. Agent will call him back at the same number as before. Customer has a problem with his DSL modem. The problem is with the gray phone cord at the back part of the modem, which is behind his desk. Agent helps him to fix it over the phone.	6.86279	Model created with help from SigOpt

Figure 6. Transcript files with summary example.

Summary: SigOpt's intelligent experimentation value for LTIMindtree's future use cases

1. **Sample efficiency:** A key component of building out deep learning models is optimizing for performance. When using approaches such as grid search and other less-sample-efficient optimization methods, one of the challenges often becomes narrowing down the search space to avoid applying infinite computational resources. SigOpt, on the other hand, provides a set of Bayesian and other global optimization algorithms that are specifically designed to be as sample-efficient as possible and to reduce the computational resources needed to optimize a model.

2. **Advanced experimentation features:** SigOpt offers a wide variety of advanced experimentation features that help modelers to better align their modeling objectives with business objectives. One of these features is multimetric optimization, which allows the modeler to optimize multiple competing metrics at the same time.

3. **Ease of use:** SigOpt offers an easy-to-use client library that allows modelers to easily integrate SigOpt into what they are doing today. In addition, an intuitive web dashboard experience allows the user to store artifacts, visualize results, and work closely with other members of the modeling team and other key stakeholders.

4. **Standardized modeling workflow:** Just as importantly, SigOpt standardizes the modeling workflow, which leads to better overall model performance because it allows modelers to focus on applying domain knowledge to the problem instead of investing in developing a modeling experimentation process.

By using the SigOpt Intelligent Experimentation Platform, the chatbot response inference time decreased from 16.75 seconds to 6.86 seconds.

LTIMindtree's telecommunications project objective was to reduce call agent time by optimizing their call center automated call response and summarization system. The goal was to improve the chatbot capabilities to deliver a faster, smoother, and more natural response as well as to summarize each engagement more concisely.

The SigOpt Intelligent Experimentation Platform helped LTIMindtree achieve chatbot performance improvement through an organized viewpoint with automatic optimizations to easily find the experiments that yielded the best customized results. LTIMindtree was able to easily track their chosen metrics and visualize which optimizations best achieved those metrics. This experiment helped LTIMindtree choose the most appropriate deep learning model architecture for improved chatbot performance for both responses and engagement summaries, using customized models and metrics.

The inference time it took for the chatbot to summarize a customer call before working with SigOpt was 16.75 seconds; by using SigOpt, the inference time was reduced to 6.86 seconds. This improvement reduced the chatbot response time, meaning a faster response to telecommunications customers and a more than 50 percent decrease in human agent time on routine tasks, yielding more human agent time to support customers on calls. The engagement summaries were also improved significantly by reducing the word count and smoothing the phrasing to effectively document the customer engagement.

Additionally, LTIMindtree not only delivered on improving chatbot performance but also experienced an unexpected benefit of significant CPU performance benefits, enabled by SigOpt.

Lastly, due to the measured impact of the improvements to accuracy and inference time, LTIMindtree will now offer an AI model optimization service for call centers as a product in its marketing catalog for current and future customers.

Learn more

To learn more about LTIMindtree and its work on call center automation, watch the SigOpt Summit presentation.

For more information about the SigOpt Intelligent Experimentation Platform, visit sigopt.com, or sign up for the platform for free at sigopt.com/signup and start using it today.

intel + SIGOPT + LTIMindtree

1. Bhanu Prakash Aladahallinanjappa is an accomplished program architect with over 20 years of experience in the consumer electronics, semiconductor, and automotive industries. He has extensive experience with short-range wireless networks, Linux, firmware, machine learning and deep learning. Currently, Bhanu draws on this expertise to address a variety of problems in ML/DL at LTIMindtree partnering semiconductor companies. Within LTIMindtree, he is the go-to person for the most difficult and pressing technical challenges that require data science and optimization solutions. He is passionate about fostering the next generation of technical architects, serving as a mentor and advisor to emerging engineers across these domains.
2. All values were part of experiments/evaluation.
3. "Powering Contact Center Transformation by Using Intel AI," LTIMindtree. https://builders.intel.com/docs/aibuilders/mindtree-powering-contact-center-transformation-by-using-intel-ai.pdf
4. See www.intel.com/3gen-xeon-config. Results may vary.
5. Results reported by LTIMindtree during their experimentation.

SigOpt, an Intel company, does not control or audit third-party data. You should consult other sources to evaluate accuracy.

SigOpt, an Intel company may require enabled hardware, software, or service activation.

Intel technologies may require enabled hardware, software or service activation.

No product or component can be absolutely secure.

Your costs and results may vary.

0822/CMD/PDF