HPC Consult Ticket Analysis with SambaNova

Daisy Nsibu
dnsibu@lanl.gov
Los Alamos National Laboratory
Los Alamos, New Mexico, USA

Abstract

The HPC division at Los Alamos National Laboratory (LANL) employs specialized consultants who function as a conduit between users and the computing environment. The consultants provide technical support and expertise to users by performing tasks such as troubleshooting computational issues, providing guidance on resource allocation, and helping users maximize the capabilities of LANL's computing infrastructure.

The division uses a legacy ticketing system that tracks user issues and interactions with the HPC consultants. With over 100,000 tickets and decades of interactions, there is an amazing amount of useful information within those tickets; however, the ticket system makes it difficult to extract that information. This project addresses this challenge by developing an AI-powered web application that provides comprehensive analysis of consult queue tickets.

The web application analyzes tickets with AI inference provider SambaNova to deliver fast inference on open-source large language models. This approach readily identifies user sentiment and ticket trends, determines the most recurring issues users face, and provides support through multiple functions: generating ticket summaries, predicting categories, providing resolutions, and detecting similar tickets to better support current issues. This tool not only provides insights into users but also reveals possible trends over time that identify general system issues, while also identifying possible resolutions and initial responses.

ACM Reference Format:
Daisy Nsibu. 2025. HPC Consult Ticket Analysis with SambaNova. In *Proceedings of The International Conference for High Performance Computing, Networking, Storage, and Analysis (SC '25).* ACM, New York, NY, USA, 1 page.

1 Methodology and Implementation

1.1 Dataset

The data comes from LANL's HPC ticketing system Request Tracker (RT), consisting of over 100,000 tickets dating back to 2008. Data preprocessing included validation, transformation, and reduction using Python libraries Pandas and Polars. The transformation process created new ticket features: institution, network, and cluster architecture type based on ticket information, while the reduction step removed unnecessary variables.

SC '25, St. Louis, MO

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM

1.2 AI Implementation

SambaNova's platform was utilized for its high-speed inference capabilities using Llama 3.3-Instruct-70B. Model parameters were set for deterministic results, and few-shot prompting was designed for tasks including: summarizing tickets and resolutions, predicting ticket categories, generating HPC-related categories, justifying predictions, detecting user sentiment, uncovering users' main issues, and identifying anomalies. LANL's clusters and consultant-defined categories were used as input in the prompts, with ticket data as input in the user message.

1.3 Web Application Development

The web application was developed using Python, Streamlit for UI, Plotly for interactive visualizations, and Polars for data processing. The application includes comparative analysis, summary view, time analysis, ticket activity, and anomaly detection pages, each designed with consultant workflows in mind to support efficient ticket analysis.

2 Results and Discussion

Processing the large dataset presented significant challenges. Real-time inferencing initially overwhelmed SambaNova's compute resources, necessitating a switch to batch inferencing during off-peak hours. This approach introduced limitations, as tasks like anomaly detection could only be applied to batches rather than the entire dataset, requiring manual aggregation of results.

The LLM occasionally produced hallucinations in sentiment analysis and category prediction. To address this, a post-processing pipeline was developed to flag and remove hallucinated results, followed by reprocessing the affected data. For HPC-related category generation, the LLM created unique categories for almost every ticket, which was not the desired outcome.

3 Conclusion

We have developed the first of its kind, AI-powered web application that HPC consultants can use to easily explore and analyze ticket data. Future work includes using more advanced LLMs with greater context windows and faster processing speeds, implementing advanced prompt engineering techniques, developing hallucination guardrails, and exploring non-LLM methods for sentiment analysis, anomaly detection, and issue identification. The ultimate goal is to integrate live tickets from RT into this web application in real-time, creating significant efficiency gains by allowing consultants to focus on more complex issues.

As HPC centers regularly transition between emerging and retiring clusters, this tool provides consultants with new ways to explore ticket data, enabling more efficient knowledge extraction and application across evolving HPC landscapes. *LA-UR-25-28797*