HPC Consult Ticket Analysis with SambaNova

Author: Daisy Nsibu ; Mentors: Mike Mason, Tolulope Olatunbosun | HPC-SYS

Overview

Background

The High Performance Computing (HPC) Division uses RT, an older 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 old ticket system makes it difficult to get that information out. The consultants want to know how this underutilized data can be used to better understand and serve their users.

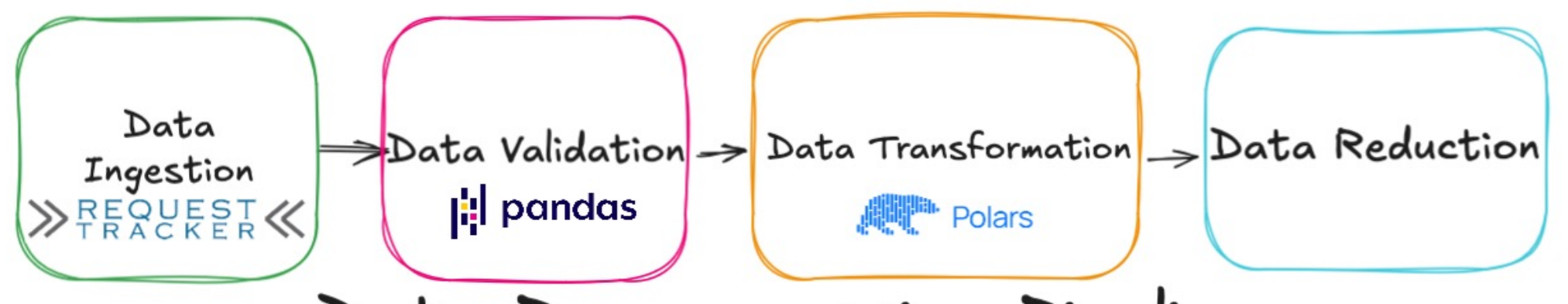
Objectives

- Provide a better method to searching for tickets
- Develop a tool that uses AI/ML that provides comprehensive analysis of consult queue tickets.
- Use LLMs hosted by SambaNova to analyze the full ticket correspondence efficiently

Methods

Data Acquisition & Preprocessing

- Extracted over 100,00 tickets from 2008 to present day
- Source: Request Tracker (RT) ticketing system



Data Pre-processing Pipeline

AI Implementation

- Utilized SambaNova platform, powered by Reconfigurable Dataflow Units (RDUs), for high-speed inference capabilities
- Implemented **Llama 3.3-Instruct-70B** model to:
  - Generate summaries
  - Predict and generate tickets categories
  - Perform sentiment and ticket analysis

Web Application Development

- Built interactive dashboard using **Streamlit** framework
- Integrated **Plotly** for advanced interactive visualizations

Conclusion

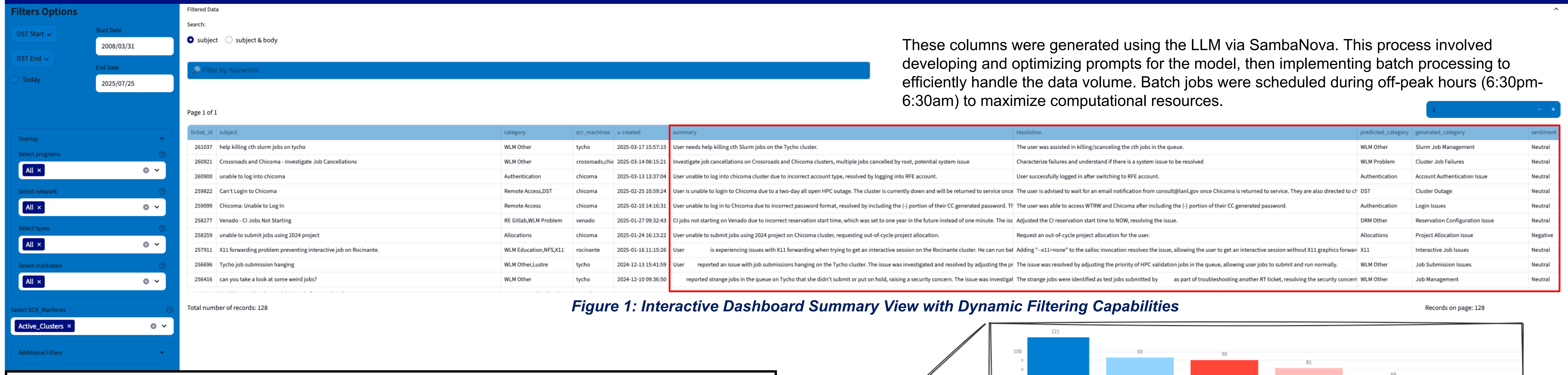
Conclusion

With over 100,000 tickets and decades of interactions collected, we were able to develop the first of its kind, AI powered web application that consultants can use to easily search for tickets as well as explore the data.

Future Directions

- Process ticket data using an advanced LLM:Llama 4 Maverick
- Live Ticket Integration: The web app connects to the RT API to collect real-time tickets
- Feedback ticket summaries, sentiment and predicted category to RT

Results



Page 1 of 1

ticket_id	subject	category	scr_machines	↓created	summary	resolution	predicted_category	generated_category	sentiment
261037	help killing cth slurm jobs on tycho	WLM Other	tycho	2025-03-17 15:57:15	User needs help killing cth Slurm jobs on the Tycho cluster.	The user was assisted in killing/scanceling the cth jobs in the queue.	WLM Other	Slurm Job Management	Neutral
260921	Crossroads and Chicoma - Investigate Job Cancellations	WLM Other	crossroads,chic	2025-03-14 08:15:21	Investigate job cancellations on Crossroads and Chicoma clusters, multiple jobs cancelled by root, potential system issue	Characterize failures and understand if there is a system issue to be resolved	WLM Problem	Cluster Job Failures	Neutral
260900	unable to log into chicoma	Authentication	chicoma	2025-03-13 13:37:04	User unable to log into chicoma cluster due to incorrect account type, resolved by logging into RFE account.	User successfully logged in after switching to RFE account.	Authentication	Account Authentication Issue	Neutral
259822	Can't Login to Chicoma	Remote Access,DST	chicoma	2025-02-25 10:59:24	User is unable to login to Chicoma due to a two-day all open HPC outage. The cluster is currently down and will be returned to service once	The user is advised to wait for an email notification from consult@lanl.gov once Chicoma is returned to service. They are also directed to ch	DST	Cluster Outage	Neutral
259099	Chicoma: Unable to Log In	Remote Access	chicoma	2025-02-10 14:16:31	User unable to log in to Chicoma due to incorrect password format, resolved by including the (-) portion of their CC generated password. Th	The user was able to access WTRW and Chicoma after including the (-) portion of their CC generated password.	Authentication	Login Issues	Neutral
258277	Venado - CI Jobs Not Starting	RE Gitlab,WLM Problem	venado	2025-01-27 09:32:43	CI jobs not starting on Venado due to incorrect reservation start time, which was set to one year in the future instead of one minute. The iss	Adjusted the CI reservation start time to NOW, resolving the issue.	DRM Other	Reservation Configuration Issue	Neutral
258259	unable to submit jobs using 2024 project	Allocations	chicoma	2025-01-24 16:13:22	User unable to submit jobs using 2024 project on Chicoma cluster, requesting out-of-cycle project allocation.	Request an out-of-cycle project allocation for the user.	Allocations	Project Allocation Issue	Negative
257911	X11 forwarding problem preventing interactive job on Rocinante.	WLM Education,NFS,X11	rocinante	2025-01-16 11:15:26	User is experiencing issues with X11 forwarding when trying to get an interactive session on the Rocinante cluster. He can run bat	Adding "--x11=none" to the salloc invocation resolves the issue, allowing the user to get an interactive session without X11 graphics forwar	X11	Interactive Job Issues	Neutral
256696	Tycho job submission hanging	WLM Other,Lustre	tycho	2024-12-13 15:41:59	User reported an issue with job submissions hanging on the Tycho cluster. The issue was investigated and resolved by adjusting the pr	The issue was resolved by adjusting the priority of HPC validation jobs in the queue, allowing user jobs to submit and run normally.	WLM Other	Job Submission Issues	Neutral
256416	can you take a look at some weird jobs?	WLM Other	tycho	2024-12-10 09:36:50	reported strange jobs in the queue on Tycho that she didn't submit or put on hold, raising a security concern. The issue was investigat	The strange jobs were identified as test jobs submitted by as part of troubleshooting another RT ticket, resolving the security concern	WLM Other	Job Management	Neutral

Total number of records: 128

Records on page: 128

These columns were generated using the LLM via SambaNova. This process involved developing and optimizing prompts for the model, then implementing batch processing to efficiently handle the data volume. Batch jobs were scheduled during off-peak hours (6:30pm-6:30am) to maximize computational resources.

Figure 1: Interactive Dashboard Summary View with Dynamic Filtering Capabilities

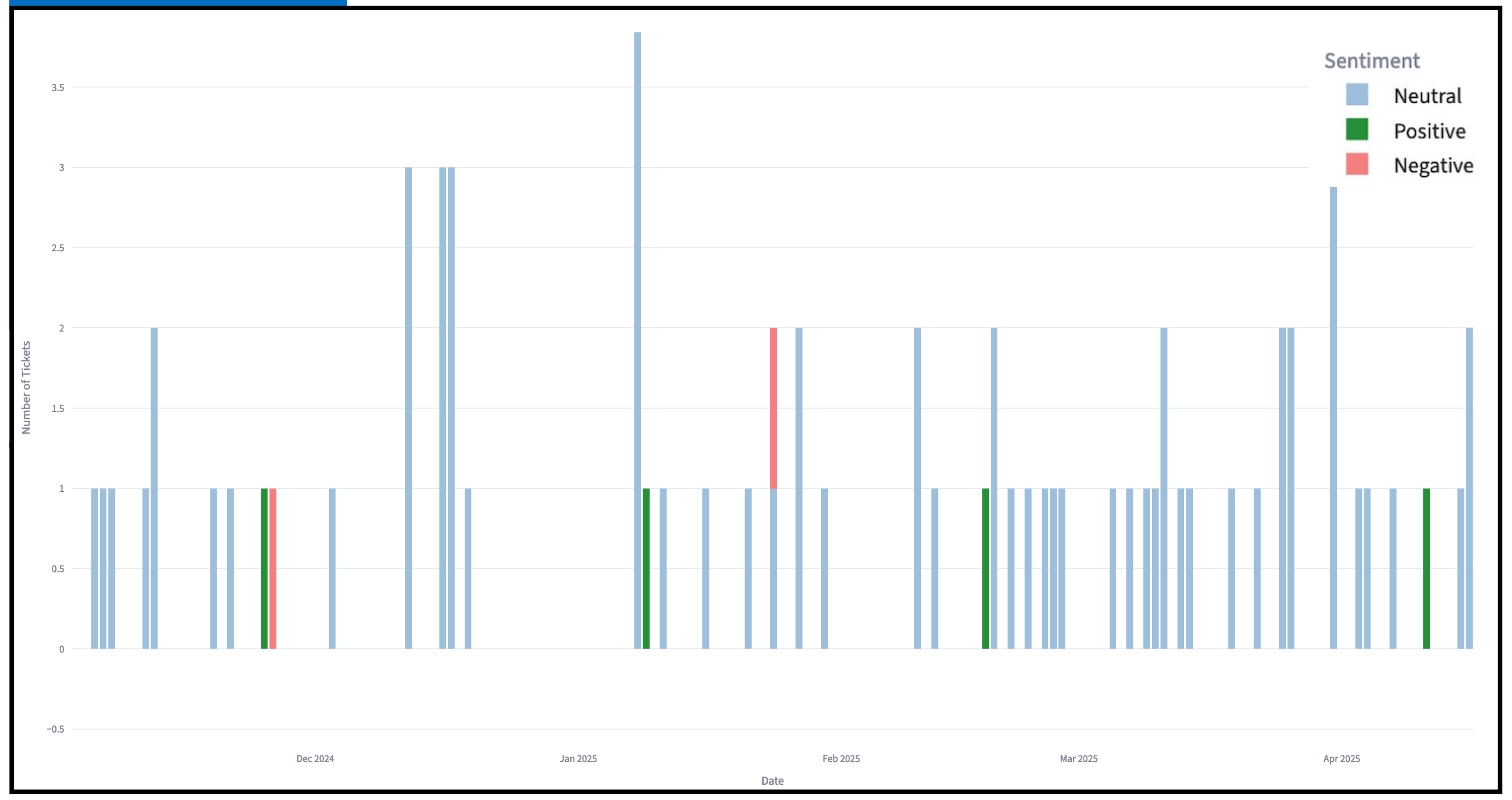


Figure 2. Sentiment Over Time View of Filtered Data

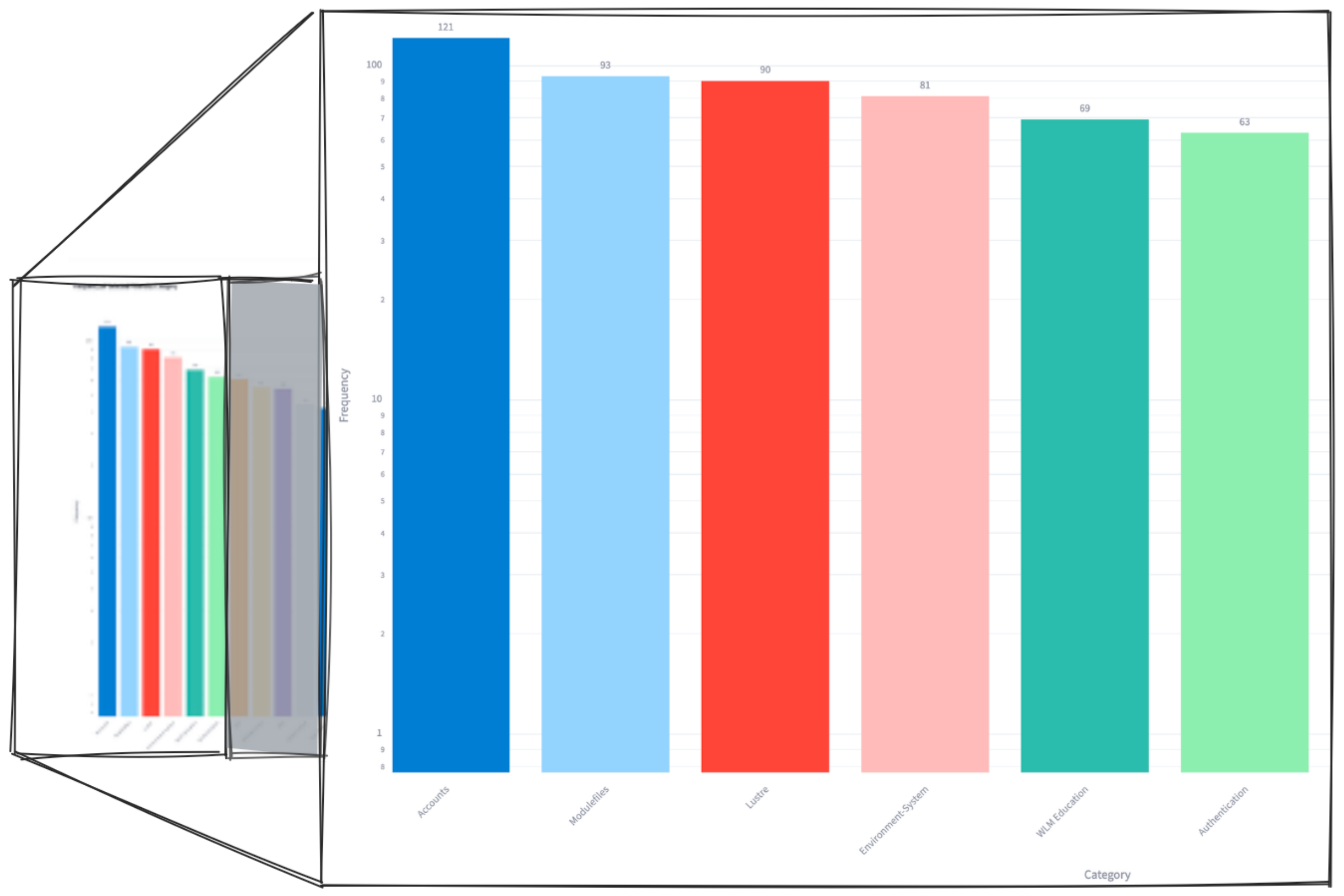
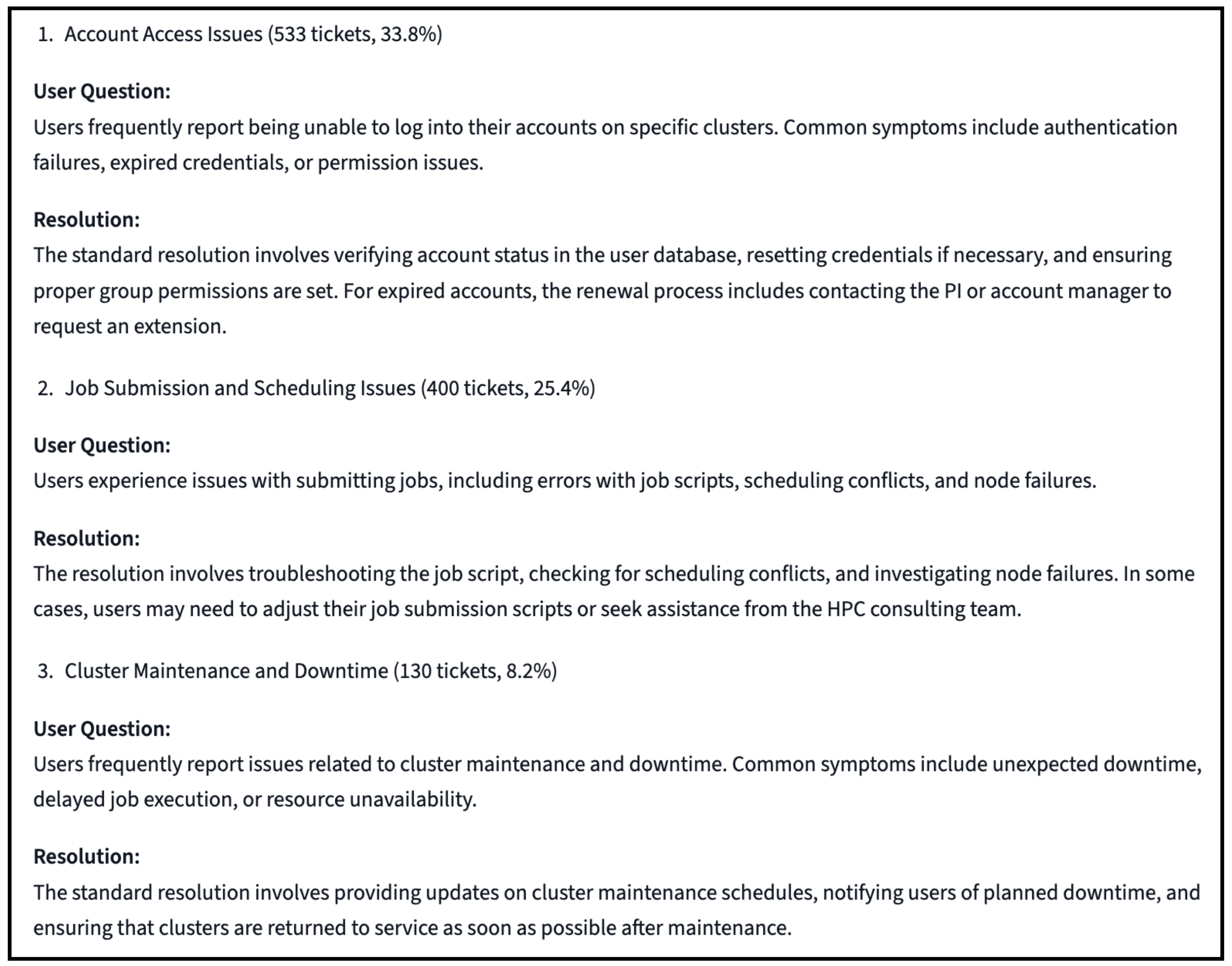


Figure 3. Zoomed in View of the Distribution of Ticket Categories

RANK	CATEGORY	LLM-PREDICTED CATEGORY	LLM-GENERATED CATEGORY
1	ACCOUNTS	SOFTWARE RELATED	CLUSTER MAINTENANCE
2	MODULEFILES	AUTHENTICATION	PROJECT CREATION AND MANAGEMENT
3	LUSTRE	ACCOUNTS	STORAGE-OTHER
4	ENVIRONMENT- SYSTEM	ALLOCATIONS	SCHEDULED MAINTENANCE
5	WLM EDUCATION	LUSTRE	PROJECT CREATION
6	AUTHENTICATION	DST	USER ACCOUNT MANAGEMENT

Table 1. Top 6 Consultant-Defined Categories with LLM-Predicted and LLM-Generated Categories



1. Account Access Issues (533 tickets, 33.8%)

User Question:
Users frequently report being unable to log into their accounts on specific clusters. Common symptoms include authentication failures, expired credentials, or permission issues.

Resolution:
The standard resolution involves verifying account status in the user database, resetting credentials if necessary, and ensuring proper group permissions are set. For expired accounts, the renewal process includes contacting the PI or account manager to request an extension.

2. Job Submission and Scheduling Issues (400 tickets, 25.4%)

User Question:
Users experience issues with submitting jobs, including errors with job scripts, scheduling conflicts, and node failures.

Resolution:
The resolution involves troubleshooting the job script, checking for scheduling conflicts, and investigating node failures. In some cases, users may need to adjust their job submission scripts or seek assistance from the HPC consulting team.

3. Cluster Maintenance and Downtime (130 tickets, 8.2%)

User Question:
Users frequently report issues related to cluster maintenance and downtime. Common symptoms include unexpected downtime, delayed job execution, or resource unavailability.

Resolution:
The standard resolution involves providing updates on cluster maintenance schedules, notifying users of planned downtime, and ensuring that clusters are returned to service as soon as possible after maintenance.

Figure 5. Top 3 User Issues and Resolution for Filtered Data

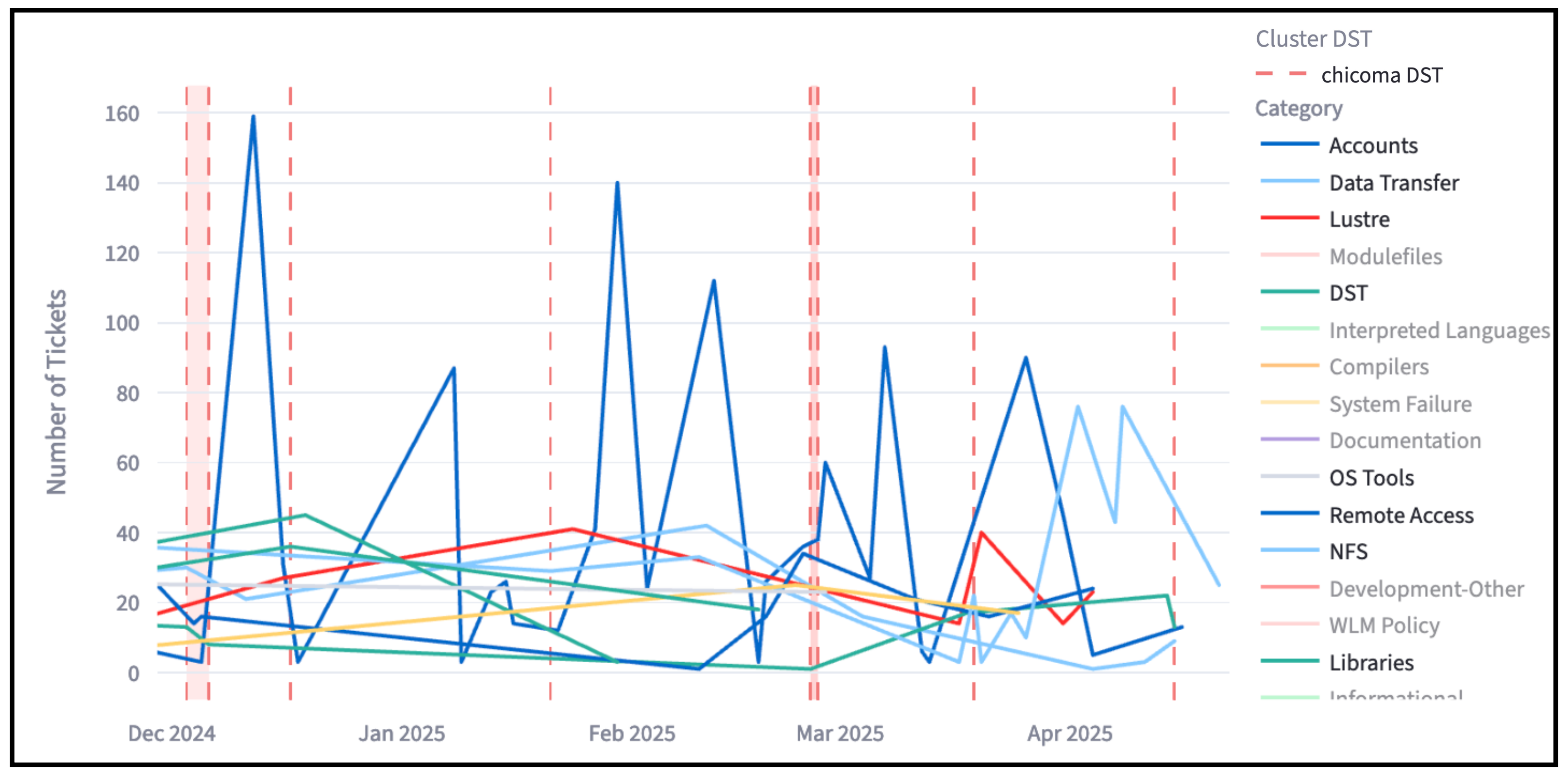


Figure 4. Number of Tickets Created by Date and Category with Chicoma DST Overlay

Los Alamos NATIONAL LABORATORY

Managed by Triad National Security, LLC, for the U.S. Department of Energy's NNSA.

LA-UR-25-27682