Decoding Windows Events: Making Logs Human-Readable with Large Language Models

Jake Ludlow, Nicholas Ramirez-Ornelas, Byron Wong, Christopher Avakian

Department of Computer Science
California Polytechnic University, Pomona

ABSTRACT

Windows Event Logs generate gigabytes of cryptic log entries daily, which can be challenging for cybersecurity professionals to interpret. This project leverages the natural language processing capabilities of Large Language Models (LLMs) to translate these logs into clear, human-readable explanations. The goal is to enhance log accessibility for a wider range of users, including IT staff, security analysts, and non-technical users such as stakeholders. By bridging the gap between raw log data and understandable insights, this project aims to improve incident response, training, and overall system comprehension.

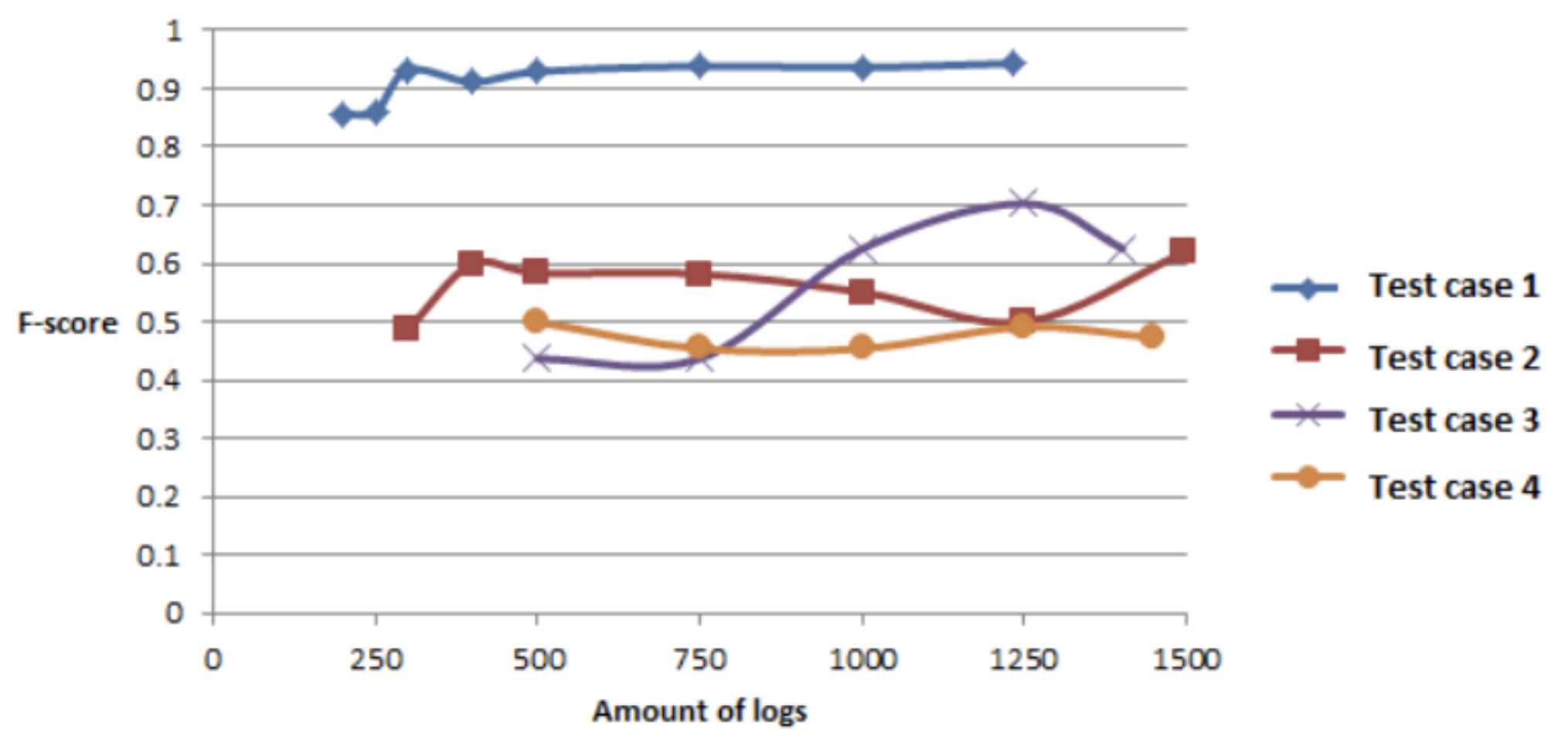

Figure 1: F-Score of different test cases of SOFM-DBSCAN Log Analysis with preprocessed data, TF-IDF, and timestamp statistics

OBJECTIVES

The purpose of this project is to develop a solution that simplifies the interpretation of Windows Event Logs, enabling faster and more efficient responses to security incidents. These event logs are often delivered in a convoluted mess of words and numbers strung together to complete a single event. When investigating a security threat towards a system, there are often hundreds or even thousands of these events. It takes precious time to identify what these events mean, conclude on the current attack, and figure out the next steps to take to mitigate the attack. Our project aims to train an LLM to take in large amounts of log data and give back a simple yet detailed conclusion on the security event and risks at hand indicated by the numerous logs.

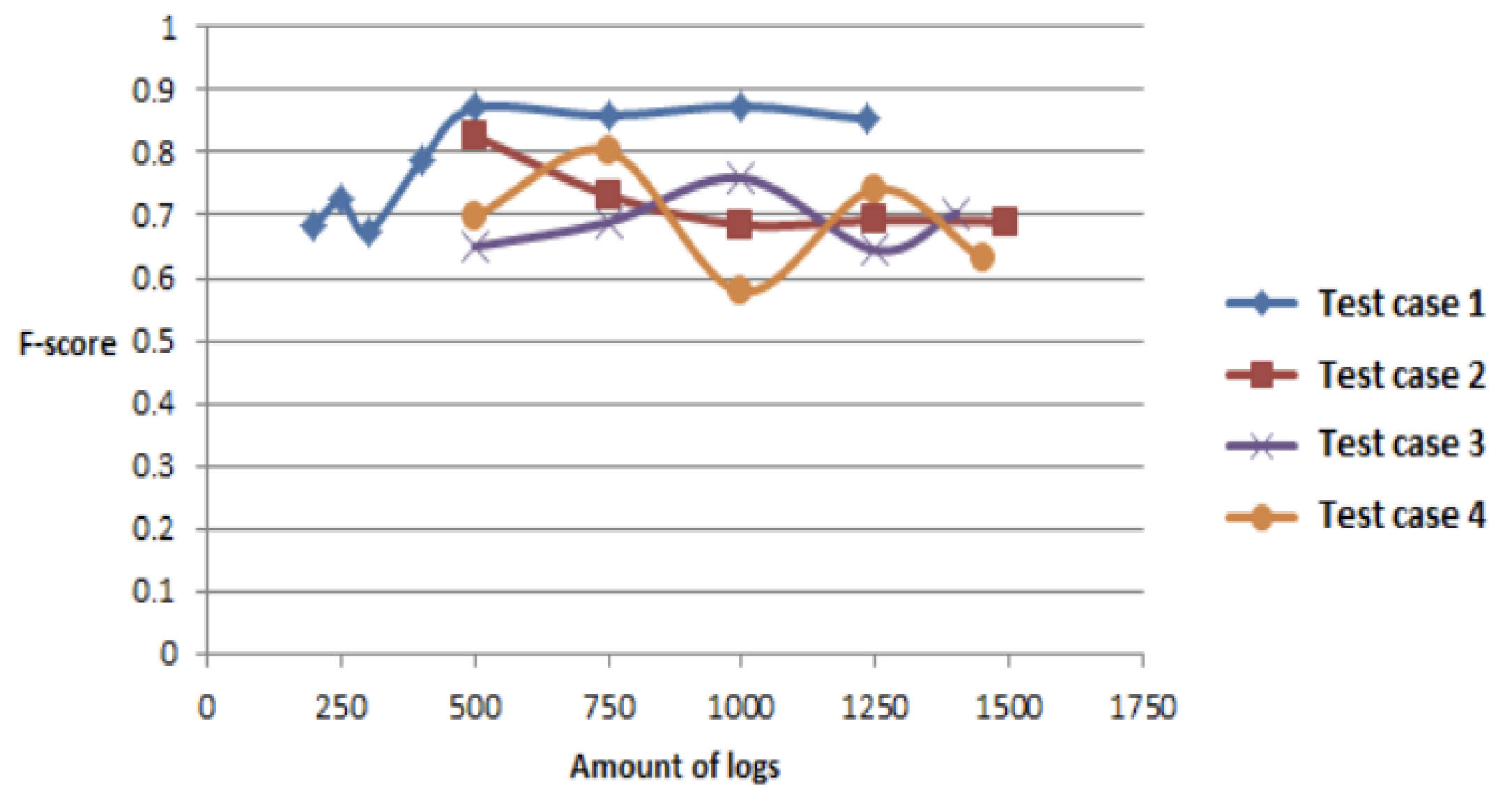

Figure 2: F-Score of different test cases of SOFM-DBSCAN Log Analysis

METHODOLOGY

The process begins by collecting a dataset of Windows Event Logs from publicly available sources to train the language model. A pre-trained language model (LLM) suited for log interpretation is then selected, followed by fine-tuning the model on specific log data to enhance its understanding of Windows Event terminology and structure. The model's performance is evaluated using a combination of qualitative feedback from IT professionals and quantitative metrics like accuracy in log interpretation to ensure its effectiveness.

Model	GA	PA	FGA	FTA	GGD
Drain	87.2	40.0	75.1	34.4	9.00
Brain	**96.6**	40.4	90.8	42.7	3.35
LogPPT	92.3	**86.5**	89.2	69.5	6.25
GPT-3.5	91.5	68.4	86.0	64.7	5.88
GPT-3.5+ICL+VA	89.8	67.2	86.1	64.8	5.81
GPT-3.5+ICL+VA+Merge	90.1	61.7	86.9	59.7	5.50
GPT-4	92.5	75.6	91.6	**75.7**	3.69
GPT-4+ICL+VA+Merge	91.8	78.5	**92.2**		

Table 1. Comparisons of LogParser-LLM on Loghub-2k via LLMs and traditional methods.

RESULTS

The results of deep learning models for log-based anomaly detection revealed some significant performance variations across datasets and experimental settings. Some notable examples being DeepLog and LogAnomaly achieved an F1-Score about 0.95 on HDFS, but these results were sensitive to data leakage when using random training data selection. Varying log grouping strategies and window sizes yielded inconsistent performance, highlighting the challenge of capturing anomalies spanning multiple sequences. The highly imbalanced nature of anomaly data poses difficulties, with models struggling to detect rare events occurring in less than 1% of the data. Introducing mislabeled logs and log parsing errors further downgraded performance, emphasizing the importance of data quality and accurate preprocessing. Recent work, however, has demonstrated the potential of Large Language Models (LLMs), specifically DistilRoBERTa, achieving a higher F1-Score than its current contemporaries suggesting a new avenue for robust and interpretable log analysis.

Accuracy of Log Interpretation: Measures how correctly the LLM translates log entries compared to expert human interpretations.

Improved Incident Response: Demonstrates a reduction in time taken to respond to system events by translating logs more quickly.

User Satisfaction: Feedback from IT professionals and non-technical users showing how accessible and helpful the translated logs are.

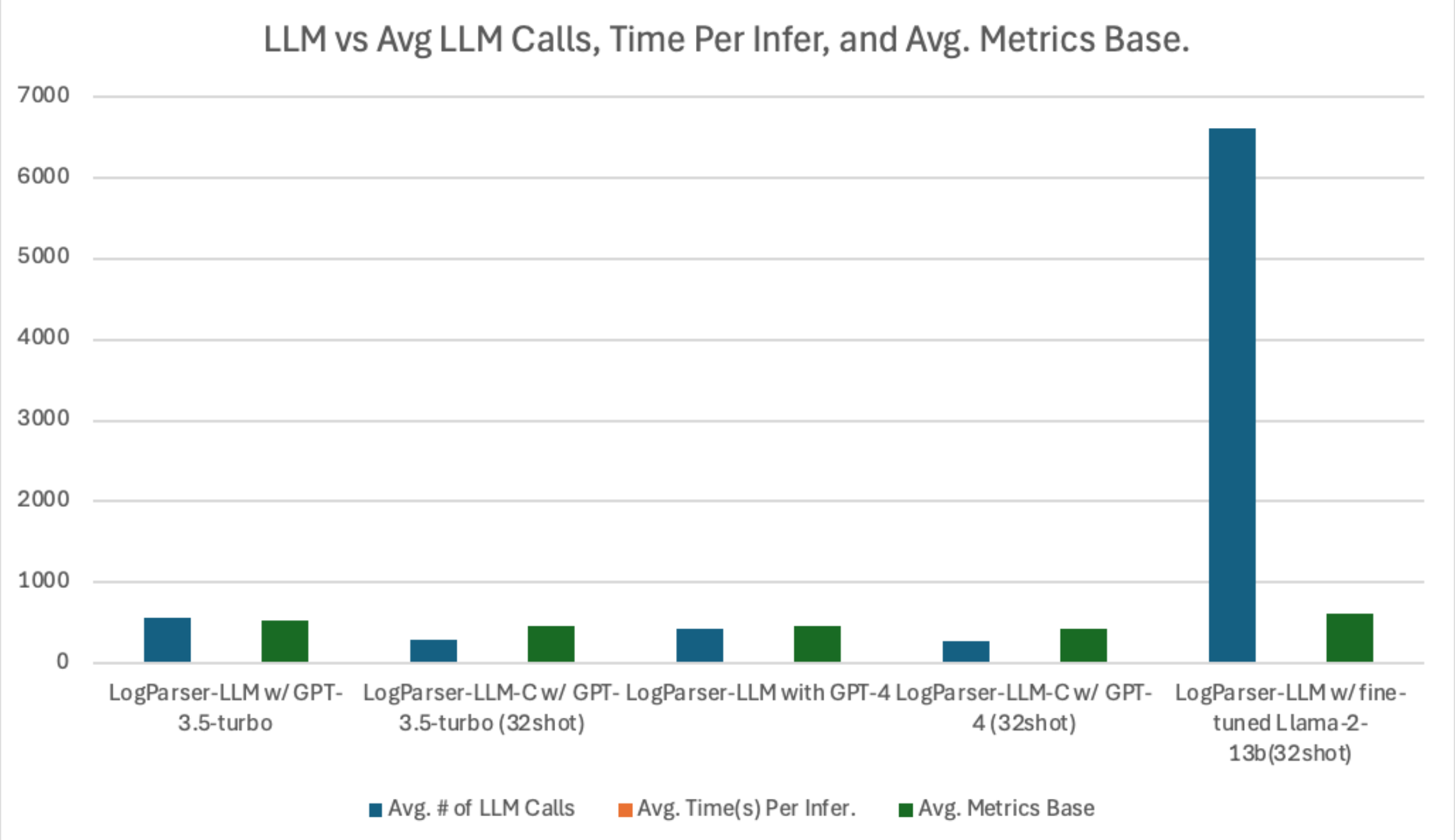

Figure 3: Efficiency and effectiveness of the LogParser-LLM with different LLMs.

CONCLUSIONS

This project highlights the potential of LLMs in bridging the gap between technical and non-technical users when it comes to system log interpretation. With the increasing complexity of security logs, the ability to translate them into plain language could significantly improve organizational incident reporting, alerting, and IT training processes. There are challenges to ensuring accuracy and consistency in translation, but the application of LLMs opens a new avenue for simplifying cybersecurity tasks.

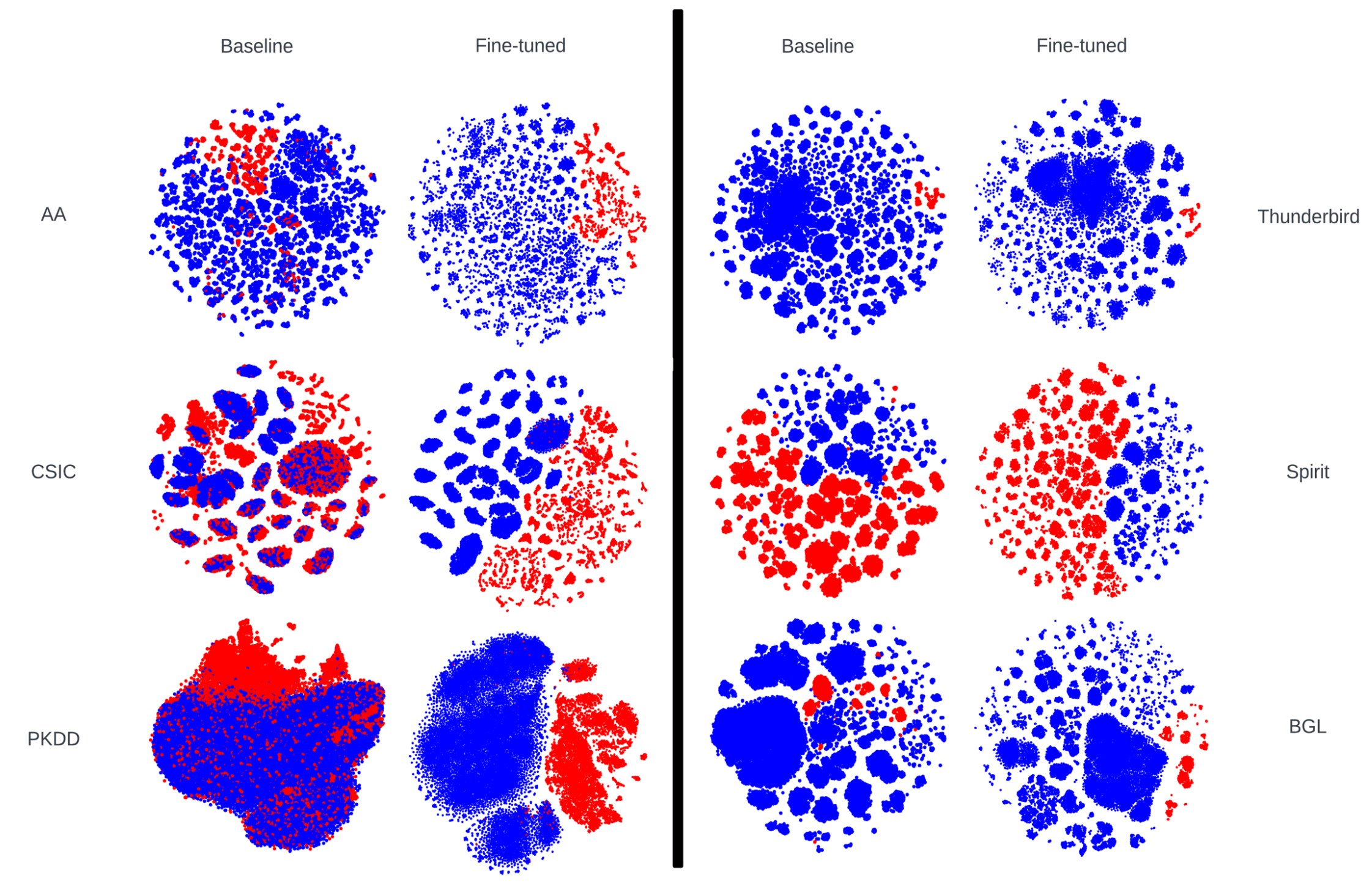

Figure 4: t-SNE visualizations of the embeddings produced by baseline and fine-tuned DistilRoBERTa models for all six datasets. Left (Right) pairs represent Application (System) logs

REFERENCES

Awotipe, O. (2019). Log analysis in cyber threat detection (Master's creative component, Iowa State University). Iowa State University Digital Repository. https://dr.lib.iastate.edu/server/api/core/bitstreams/8dd82e7d-2e78-4e2a-9bc9-cc9bc68c1fe0/content

Bareiss, P., & Hernandez, J. (2023). attack_data [Source code]. GitHub. https://github.com/splunk/attack_data

Guo, H., Lin, X., Yang, J., Zhuang, Y., Bai, J., Zheng, T., Zhang, B., & Li, Z. (2022). TransLog: A unified transformer-based framework for log anomaly detection. arXiv. https://arxiv.org/abs/2201.00016

Karlsen, E., Luo, X., Zincir-Heywood, N., & Heywood, M. (2023). Benchmarking large language models for log analysis, security, and interpretation. arXiv. https://arxiv.org/abs/2311.14519

Le, V.-H., & Zhang, H. (2022, May). Log-based anomaly detection with deep learning: How far are we? In Proceedings of the 44th International Conference on Software Engineering (ICSE '22). ACM. https://doi.org/10.1145/3510003.3510155

Li, W. (2013). Automatic log analysis using machine learning: Awesome automatic log analysis version 2.0. https://uu.diva-portal.org/smash/get/diva2:667650/FULLTEXT01.pdf

Liu, Y., Tao, S., Meng, W., Wang, J., Ma, W., Zhao, Y., Chen, Y., Yang, H., Jiang, Y., & Chen, X. (2024). Interpretable online log analysis using large language models with prompt strategies. ArXiv. https://arxiv.org/abs/2308.07610

Zhang, C., Song, D., Chen, Y., Feng, X., Lumezanu, C., Cheng, W., Ni, J., Zong, B., Chen, H., & Chawla, N. V. (2019). A deep neural network for unsupervised anomaly detection and diagnosis in multivariate time series data.
Proceedings of the AAAI Conference on Artificial Intelligence, 33(01), 1409-1416. https://doi.org/10.1609/aaai.v33i01.33011409

Zhong, A., Mo, D., Liu, G., Liu, J., Lu, Q., Zhou, Q., Wu, J., Li, Q., & Wen, Q. (2024). LogParser-LLM: Advancing efficient log parsing with large language models. arXiv. https://arxiv.org/abs/2408.13727

ACKNOWLEDGEMENTS

We would like to thank **Professor Amamra,** and the **California Polytechnic University, Pomona.**

RESEARCH POSTER PRESENTATION DESIGN © 2019
www.PosterPresentations.com