Spectra AI Mini Challenge: Anomaly Prompt Detection

Aditya Pawar

Deliverables:
- Spectra_AI_Anomaly_Prompt_Detection (Natural Language Prompts).ipynb
- Spectra_AI_Anomaly_Prompt_Detection (Synthetic Embeddings).ipynb
- prompts_normal.txt
- prompts_anomalous.txt

1. Problem Understanding and Rationale

Language models (LLMs) are vulnerable to anomalous or malicious user inputs, such as jailbreak or system-prompt injection attempts.
These adversarial prompts can manipulate the model into revealing sensitive data or executing unintended instructions.

The objective of this project is to design a lightweight prototype that can detect anomalous or malicious prompts based on their high-dimensional embeddings,
using mathematical and statistical techniques rather than large neural networks.

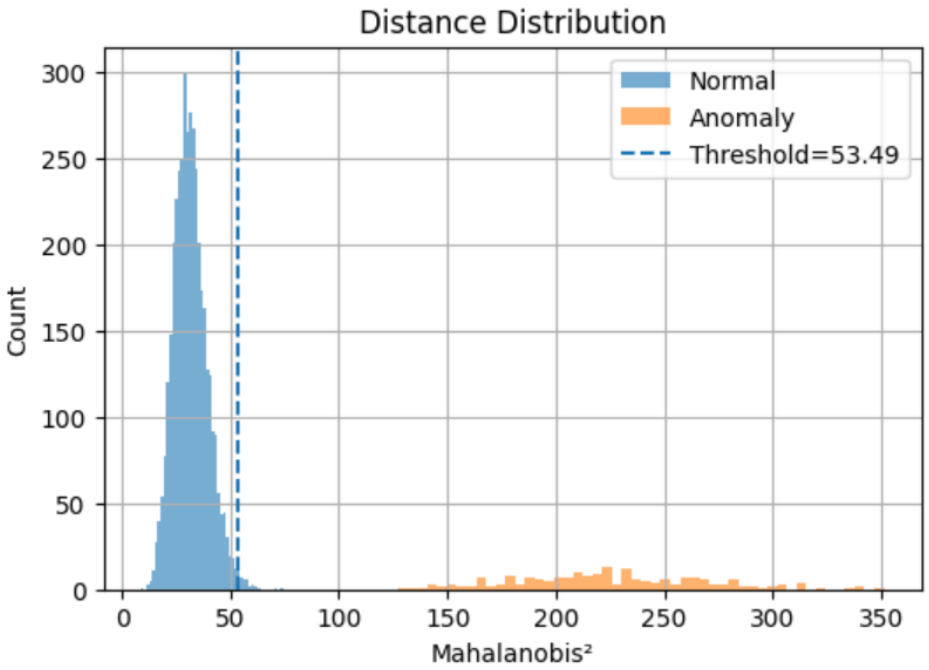
Two complementary approaches are implemented:
1. Natural Language Prompt Embeddings using SentenceTransformer ('all-MiniLM-L6-v2').
2. Synthetic Embeddings Simulation using multivariate Gaussian vectors.

2. Design and Implementation

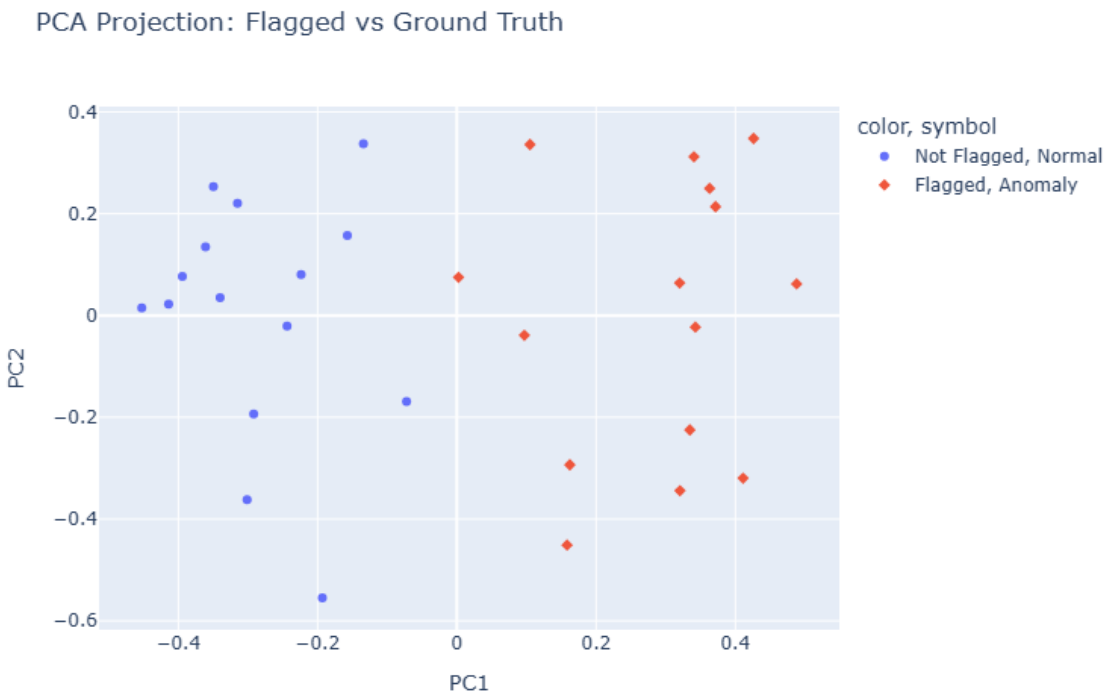
1. Data Preparation: Prompts are loaded and embedded into high-dimensional vectors.
2. Linear Algebra: Compute covariance matrix and Mahalanobis distance for anomaly scoring.
3. Probability: Apply χ^2 thresholding to identify anomalies statistically.
4. Bayesian Analysis: Compute posterior probability of true anomalies given prior rates.
5. Explainable AI (xAI): Decompose Mahalanobis^2 into per-dimension contributions.
6. Visualization: Histogram, PCA, t-SNE, QQ-plot, ECDF, Threshold sweep.

3. Results and Observations

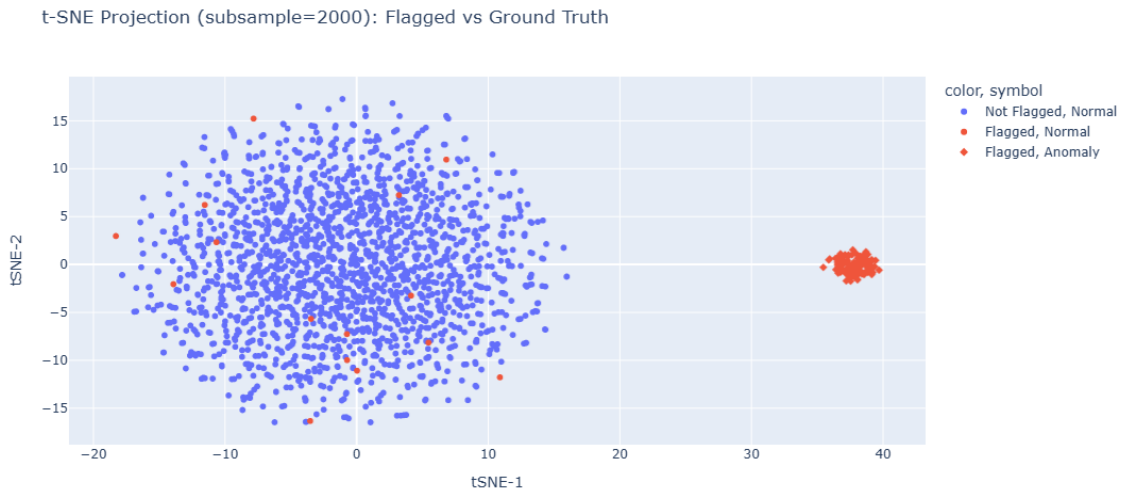
Statistical separation between normal and anomalous prompts was clearly visible in all visualizations.



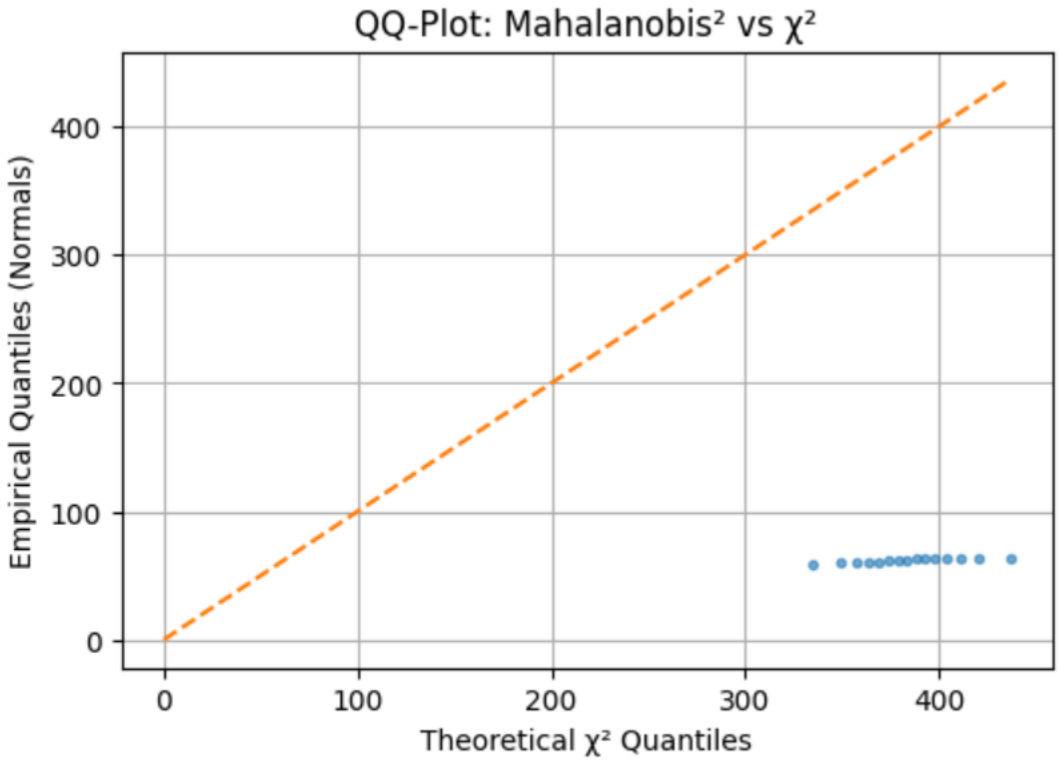
Histogram of Mahalanobis^2 for Normal vs. Anomalous Prompts



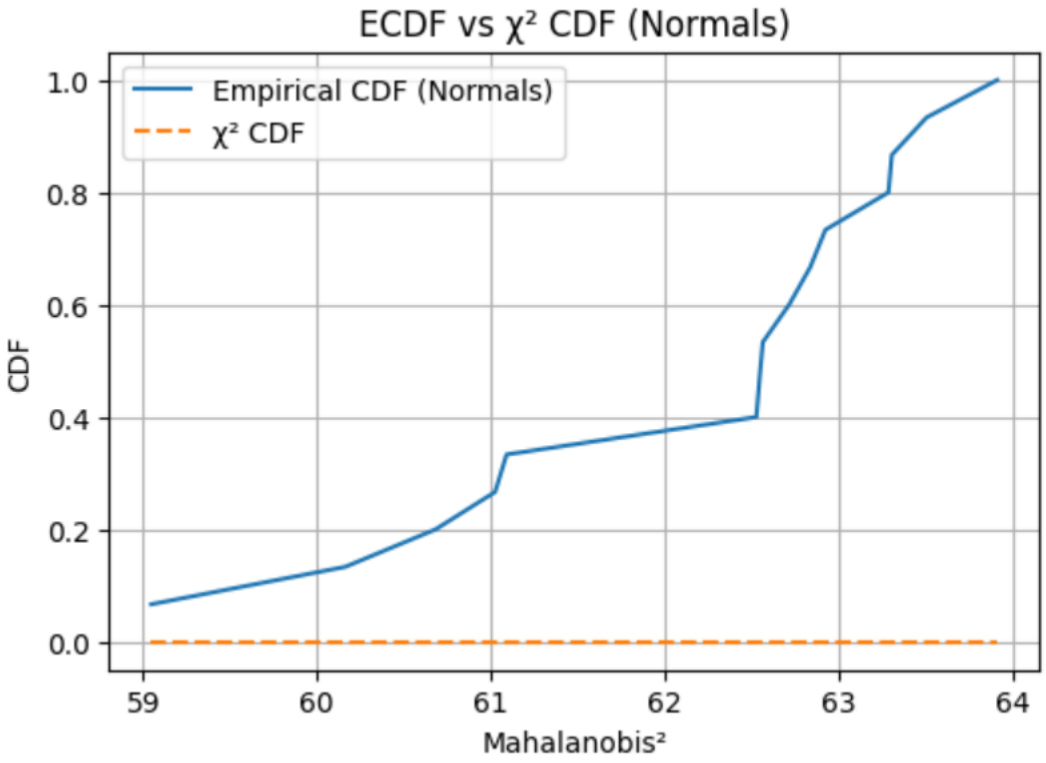
PCA 2D Scatter Plot



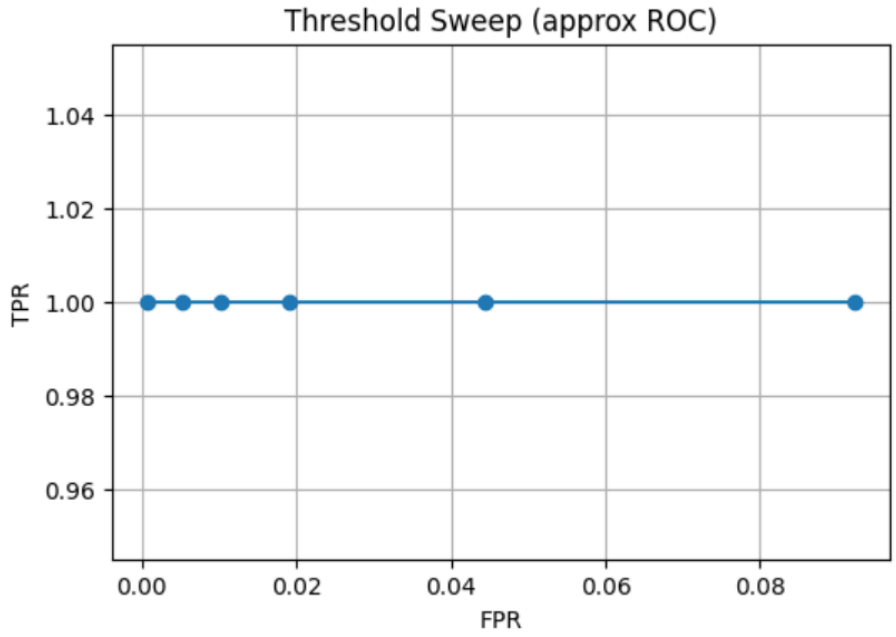
t-SNE Plot (Flagged vs. Ground Truth)



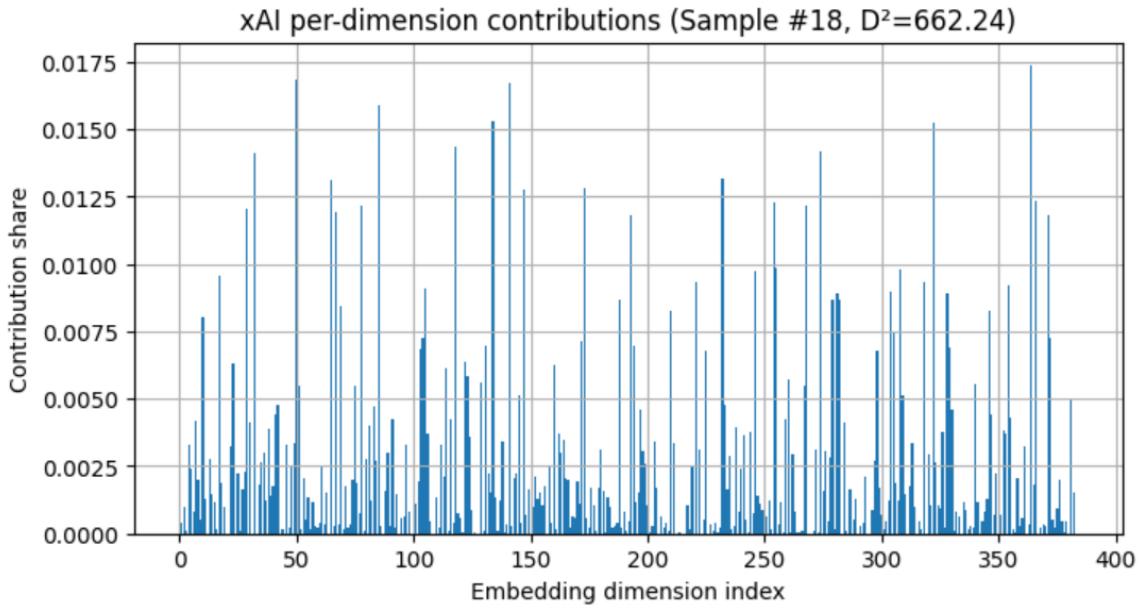
QQ-Plot Comparing Mahalanobis^2 vs. χ^2



ECDF vs. χ^2 CDF Plot



Threshold Sweep



Bar Chart of Per-Dimension Contributions

The histogram shows clear separation between normal and anomalous prompts, and the PCA/t-SNE plots indicate distinct clusters.
The QQ and ECDF plots validate the χ^2 assumption. Bayesian PPV analysis (prior anomaly rate = 2%) yields a posterior of around 0.39, meaning 39% of flagged prompts are truly malicious under realistic conditions.

4. Security, Ethical, and Governance Considerations

- False Positives: Overly strict thresholds may flag benign prompts; calibration required.
- Privacy: Only embeddings are analyzed, preserving user confidentiality.
- Model Drift: Covariance should be retrained periodically to maintain accuracy.
- Governance: Integrate as part of multi-layered safety system with logging and audit trails.

5. Comparative Summary

Aspect	Synthetic Embeddings	Natural Language Prompts
Data Source	Simulated Gaussian vectors	Real text via SentenceTransformer
Dimensionality	Configurable (8–64)	Fixed (384)
Speed	Very fast	Slightly slower (embedding overhead)

Realism	Theoretical	Practical and deployment-ready
Use Case	Mathematical validation	Real-world monitoring prototype

6. Conclusion

This project demonstrates how linear algebra and probability can form the backbone of an interpretable, data-efficient anomaly detector for AI systems.
Combining Mahalanobis distance, χ^2 hypothesis testing, Bayesian reasoning, and xAI interpretation results in a statistically rigorous yet transparent anomaly detection prototype.

This system can be extended to production AI governance pipelines for detecting anomalous prompts, enhancing security, accountability, and trustworthiness in AI deployments.

Attachments

- Spectra_AI_Anomaly_Prompt_Detection (Natural Language Prompts).ipynb
- Spectra_AI_Anomaly_Prompt_Detection (Synthetic Embeddings).ipynb
- prompts_normal.txt, prompts_anomalous.txt
- Visual Figures (PCA, t-SNE, QQ, ECDF, Histogram, xAI, Threshold Sweep)