Using LLM Embeddings with Similarity Search for Botnet TLS Certificate Detection

University of South Florida: Kumar Shashwat, Francis Hahn, Xinming Ou

Rapid7: Stuart Millar

PROBLEM

- Botnets commonly use TLS to connect with C&C servers.
- Features of TLS certificates could indicate being used by botnet activity.
- Can LLM embeddings of certificates be used for botnet detection?

TLS Fields

TLS Certificate used by AsyncRAT C&C

TLS Certificate used by a Benign Website

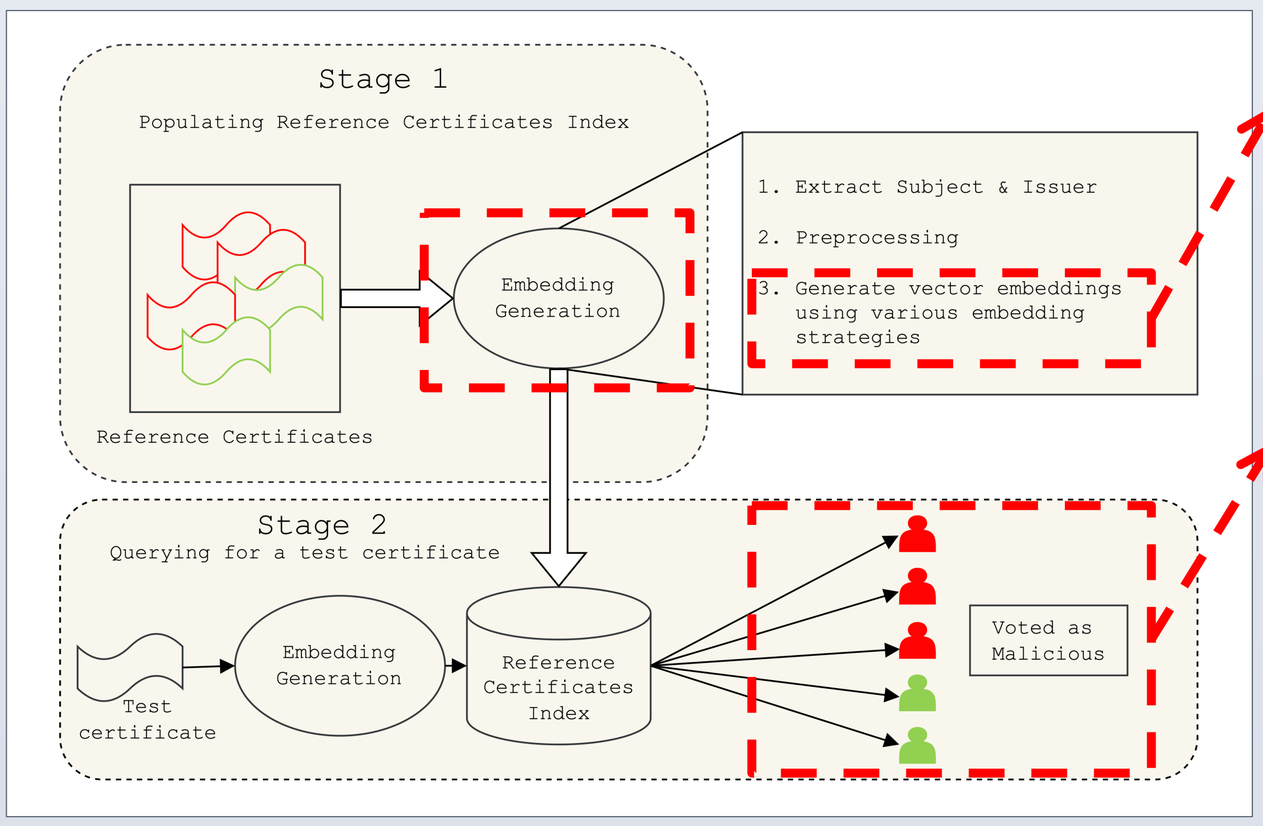
Name	Country	Organization	Organization Unit	Location	State	Email
192.236.160.249	mn	xxxyz	zaaaab	sttuvwww	noopqrrrrs	ccdde@192.236.160.249
www.agl.com.au	AU	AGL Energy Limited	NA	Docklands	Victoria	NA

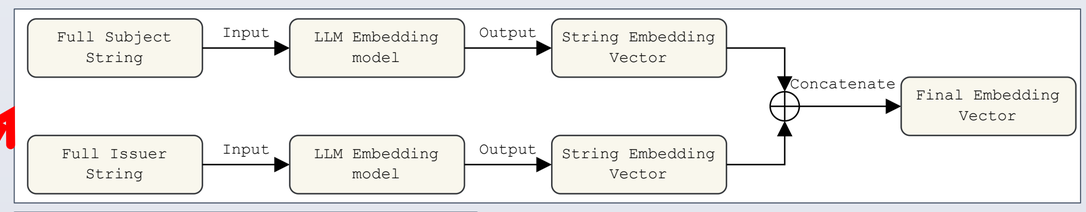
OUR CONTRIBUTION

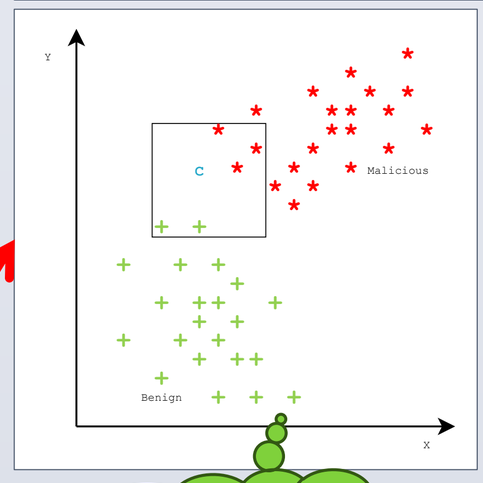
Ablation Study: Rigorous comparison of multiple proprietary and open-source LLM embedding models and selecting the best parameters and strategies.

Real World Validation: Model was evaluated over 150,000 real world certificates collected from Jan – May 2024 from Rapid7's Sonar data

RESEARCH METHODOLOGY







Majority Vote

- Lightweight and scalable
- Uses kNN and majority voting for decision making.

RESULTS

Ablation Study

Conducted ablation study to select the embedding strategy, k, and embedding model

k: 1, **5**, 10, 15, 20

Distance Metric: **COSINE**, EUCLIDEAN, Dot product, MANHATTAN

Embedding Model: **OpenAI**, BERT, Cohere, VoyageAI, Titan, Titan 2, **Character- BERT**

100-fold reduction in human effort

Real World Evaluation

- Our model flagged 13 malicious certificates out of 150,000 where 1 was confirmed malicious by VirusTotal.
- A random selection of 1,300 certs found none.

ACKNOWLEDGEMENT

This work is partially supported by the National Science Foundation under award no. 2235102, and Office of Naval Research under award no. N00014-23-1-2538. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of these parties