UofT PhishNet

(Q1) Team 11

(Q2) Abhay Singh Thakur (1011878096) | Johnny Kim (1012475297) | Darya Zanjanipour (1003605814) | Marc Bishara (996877447)

(Q3) This project aims to address the significant threat of phishing attacks targeting our university's email system. We will develop a robust machine learning model to automatically classify incoming emails as either legitimate or phishing. To achieve this, a Transformer-based model, similar to the BERT architecture used by Tida & Hsu (2022), will be trained on the labelled body text from three public datasets: one from Kaggle and two from Zenodo, including the TREC-05 phishing corpus. The model's performance will be rigorously evaluated on a held-out test set using standard classification metrics, including accuracy, precision, recall, and the F1-score.

(Q4)

- Phishing Email Detection from Kaggle by Chakraborty, S. (2023) has 18.6K rows of email content 61% of which are safe, 39% are phishing
- Phishing validation emails dataset from Zenodo by Miltchev, R. et al. (2025) has 2K rows of email content 50% of which are safe and 50% are phishing
- TREC-05.csv from the Phishing Email Curated Datasets from Zenodo by Arifa Islam, C. (2023) has 57K entries with email content of which 60% are safe and 40% are phishing

---

(Q5) This project has been previously attempted with success. For example, Xue, Spero, Koh, and Russello (2025) proposed *MultiPhishGuard*, an LLM-based multi-agent system for detecting phishing emails and reported **high accuracy**. Also, Adwan and Abuhasan (2016) developed an intelligent classification model using phishing term weighting and achieved **high accuracy** in classifying phishing vs. non-phishing emails, as well as another successful attempt by Gupta et al (2024). Given that we also have access to structured, labelled datasets, we can compare our method with these existing ones and aim to equal or surpass their performance.

(Q6) We are also in communication with UofT's Data Governance (Common Review Process) team to get access to a phishing emails dataset from the university's servers

---

References

Adwan, Y., & Abuhasan, A. (2016). *An intelligent classification model for phishing email detection*. arXiv. https://arxiv.org/abs/1608.02196

Arifa Islam, C. (2023). Phishing Email Curated Datasets [Data set]. Zenodo. https://doi.org/10.5281/zenodo.8339691

Chakraborty, S. (2023). *Phishing email detection* [Data set]. Kaggle. https://www.kaggle.com/datasets/subhajournal/phishingemails

Gupta, B. B., Gaurav, A., Arya, V., Attar, R. W., Bansal, S., Alhomoud, A., & Chui, K. T. (2024). Advanced BERT and CNN-based computational model for phishing detection in enterprise systems. *Computer Modeling in Engineering & Sciences, 0*(0), 1–10. https://doi.org/10.32604/cmes.2024.056473

Miltchev, R., Rachev, D., & Genov, E. (2025). *Phishing validation emails dataset* [Data set]. Zenodo. https://doi.org/10.5281/zenodo.13474746

Tida, V. S., & Hsu, S. (2022). *Universal spam detection using transfer learning of BERT model*. arXiv. https://arxiv.org/abs/2202.03480

Xue, Y., Spero, E., Koh, Y. S., & Russello, G. (2025). *MultiPhishGuard: An LLM-based multi-agent system for phishing email detection*. arXiv. https://arxiv.org/abs/2505.23803