National College of Ireland

Multi-Model Embedding Fusion for Phishing URL Detection: Combining GPT-3.5 Semantic Understanding with BERT Contextual Analysis for Enhanced Threat Classification

MSc Research Project
MSc in Cybersecurity

Sai Raghu Ram Raghavapurapu
Student ID: 23380195

School of Computing
National College of Ireland

Supervisor: Jawad Salahuddin

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Sai Raghu Ram Raghavapurapu
Student ID:	23380195
Programme:	MSc in Cybersecurity **Year:** 2025-26
Module:	Research Practicum
Supervisor:	Jawad Salahuddin
Submission Due Date:	28-01-2026
Project Title:	Multi-Model Embedding Fusion for Phishing URL Detection: Combining GPT-3.5 Semantic Understanding with BERT Contextual Analysis for Enhanced Threat Classification
Word Count:	8477 **Page Count** 21

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Sai Raghu Ram Raghavapurapu
Date:	28-01-2026

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Multi-Model Embedding Fusion for Phishing URL Detection: Combining GPT-3.5 Semantic Understanding with BERT Contextual Analysis for Enhanced Threat Classification

Sai Raghu Ram Raghavapurapu
23380195

Abstract

Phishing attacks remain a threat to the security of cyber-space, and existing solutions to this issue are based solely on single-model designs, either focusing on semantic aspects or structural aspects but not both simultaneously. The current work aims to introduce a new model based on a cross-attention fusion structure to bridge the gap in the current techniques of embedding in identifying phishing URLs through the utilization of the GPT-3.5 semantic embeddings and the BERT structural embeddings. The approach uses the OpenAI text-embedding-3-large model to produce embeddings of size 3,072-dimensions based on the URL descriptions, in parallel with the domain-specific 'DomURLs_BERT' model creating embeddings of size 768-dimensions based on the structural information. The two embeddings are combined in a bidirectional cross-attention architecture with an adaptive fusion gate to determine the weights for the embeddings to be fused. The results of the experimentation using the PhiUSIIL dataset with a total of 235,795 URLs show that the fusion architecture has an accuracy of 97.27%, an F1-score of 97.66%, a recall value of 99.49%, and an AUC-ROC. It has been verified that both the embedding sources provide considerable complementary information through the analysis of the fusion weights learned and calculating the contribution of BERT embeddings to be 62.71% and GPT embeddings to be 17.26%, while the cross-attention components account for the remaining 20%. The cross-attention fusion has been found to successfully leverage the complementary semantic and structural representations for enhanced threat detection in the form of a GPT-BERT model application in phishing URL detection tasks in the current work, and this is a first in its area and domain.

1 Introduction

Phishing has proved to be one of the most enduring and financially costly forms of cyber-attacks in the current cyber environment. Social engineering attacks of a sophisticated nature hoodwink victims into divulging crucial information, such as financial and personal identification details, in addition to corporate login credentials, through exploiting both the technical and psychological vulnerabilities. The Anti-Phishing Working Group has reported an unprecedented increase in phishing attacks, where more than 1.07 million phishing pages and five million attacks in total in the year 2023 have been reported by researchers (Haq et al., 2024). Contemporary phishing attacks cannot be dealt with by adopting traditional methods

considering the exponential increase in the number of attacks and the evolving complexities involved in the process.

The implications of phishing are far-reaching and include the global impacts of economic stability, privacy, and security that are compromised through phishing, where the entry point of ransomware and crucial data breaches occurs through phishing attacks (Gupta et al., 2021). The rise in the utilization of cloud services and telework has resulted in an increase in the attack surface available to malicious users, thereby creating an immediate need to develop solutions for the detection process to function efficiently in varying levels of technical settings. Additionally, the emergence of content generation through AI has allowed cyberattacks to produce phishing emails that are difficult to filter out through traditional mechanisms.

The current state-of-the-art phishing URL identification relies heavily upon single-model deep learning and traditional feature engineering settings. Machine learning solutions based upon human-designed linguistic and structural elements have managed to achieve accuracy levels ranging from 92% to 98% (Safi & Singh, 2023), while the application of transformer-based models, PhishBERT, has recorded incredible results in terms of metrics of up to 99.98% (Wang et al., 2023) for the respective benchmark datasets. Though, with the presence of adversarially altered URLs and the issue of zero-day attacks based upon the vulnerabilities of model-specific, the limitation in the above approaches appears to be intrinsic. The semantic objectives in complex phishing actions are not represented through deep learning models based solely upon structural objectives, while traditional feature engineering methods involve human modifications to adjust to the evolving pattern of phishing attacks.

The first drawback in the current techniques is their inability to analyze the structural makeup and semantic information in suspicious URLs simultaneously. While the existing BERT-based models are very effective in observing character-level and structural abnormalities, they fail to utilize the worldly information to recognize social engineering and brand spoofing. At the same time, large language models show variability in their capability to recognize structural abnormalities but perform well in semantic understanding (Mahendru & Pandit, 2024). The above-stated drawbacks suggest the need to involve a complementary approach that employs the strengths of different architectures to ensure the highest level of phishing identification efficiency.

The research question to be solved by the study is: How does the combination of GPT-3.5 semantic embeddings and BERT contextual embeddings via cross-attention improve the accuracy of phishing URL identification, particularly for sophisticated attack vectors? The question reflects a major gap in the current literature because there has been a lack of studies in combining the semantic understanding strengths of GPT-3.5 and the structural strengths of the BERT model in a study involving the security domain. Although the feasibility of the combination of GPT-BERT model has been proved in a fake news detection framework with a high level of accuracy of 95.30% (Dhiman et al., 2024), this approach has not been extended to the phishing detection domain.

The objectives of the research work that will guide the investigation include four basic objectives. The first objective involves developing a two embedding extraction model with the capability to produce GPT-3.5 semantic embeddings of 3,072 dimensions and structure embeddings of 768 dimensions from URL input values. The second objective involves developing and designing a cross-attention model to allow the flow of information in opposite directions across the embedding space. The third objective aims to accomplish the task with a detection performance of more than 97% accuracy and an F1-score higher than 0.97, thus attaining competitive results relative to the latest state-of-the-art methods for a solo model solution. The fourth objective aims to evaluate the complementarities of the multi-modal fusion solution through the investigation of the contribution weights learned for each source of the embeddings.

The methodology in the current study employs the PhiUSIIL dataset, where the number of labelled URLs amounts to 235,795. The semantic embeddings and structural embeddings will be derived by employing the OpenAI text-embedding-3-large API, and the domain-specific model DomURLs_BERT, respectively, following the URL enrichment phase. A novel cross-attention-based fusion technique with adaptive learnable weights to balance the embedding sources is employed for the combination of the diverse embeddings. Conventional classification standards, analyses of the weights in the fusion process, and comparison with popular benchmark techniques are considered in the complete testing process.

The following are the remaining sections in this report. For creating a research background and to validate the proposed approach, Section 2 presents a critical review of the literature regarding phishing detection based upon transformer-based solutions, deep learning models, hybrid solutions, and traditional machine learning methods. The methodology of the study, involving properties of the datasets, methods for learning embeddings, and experiment design, will be discussed in detail in Section 3. The design specification of the proposed fusion model will be described in Section 4. The technical implementation details will be discussed in Section 5. The results and observations will be discussed in Section 6. The summary and recommendations for the future work will be presented at the end of Section 7.

2 Related Work

Research work in phishing URL detection has grown from heuristic-based methods to complex deep learning models, but the current state-of-the-art mostly relies upon single-model architectures that are incapable of learning structural and semantic information in phishing URLs concurrently. This critical review will analyze the evolution of detection techniques based upon three categories: traditional machine learning models, transformer architecture-based models, and embedding-based models, thus forming the basis of the new GPT-3.5 and BERT-based fusion technique.

2.1 Machine Learning and Deep Learning Approaches

The classical phishing detection methods synthesised basic techniques based upon feature engineering and statistical learning. Da Silva et al. (2019) assessed heuristic-based phishing

detection for twelve URL features that include three categories: blacklist evasion, URL structure, and user susceptibility. Through their assessment of 189,892 phishing pages, they found moderate levels of relevancy scores for the feature variables, proving that feature bundles make marginal, and not dramatic, improvements in the efficiency of phishing detections. The inherent weakness in heuristic-based methods is their inability to make predictions based on unseen attack examples working to avoid the already defined detection rules.

Machine learning techniques made great strides in terms of performance gain through effective feature engineering. The importance of feature engineering in phishing detection was demonstrated by Gupta et al. (2021), where they reached a detection accuracy of 99.57% by taking only nine lexical feature variables and applying the Random Forest Classification model to the ISCX URL-2016 dataset. The importance of careful feature engineering and the effect of biased datasets on phishing detection models were explored in the study conducted by Hannousse and Yahiouche (2021), and the authors reached an accuracy of 96.61% while proposing a benchmark framework for phishing detection models. Systematic reviews involving a total of 80 publications conducted by Safi and Singh in 2023 reported Random Forest to be the classic classifier in 31 studies and PhishTank to be the primary source of phishing data in 53 studies. The review recorded a maximum accuracy rate of 99.98% achieved by Convolutional Neural Networks, but conveniently verified that none of the current studies utilize the structural assessment of BERT and semantic insight of the large language model.

The application of deep learning models facilitated the automatic extraction of features directly from the raw form of the URLs. Haq et al. (2024) designed a one-dimensional CNN model that reached a maximum accuracy of 99.7% in classifying 400,000 URLs easily and simply through their architectural design, proving the fact that complex deep learning solutions are not always the best solutions to problems. The model designed by Nanda et al. (2024) combined BiLSTM-GHA-CNN, but its computational complexities make its application in real-time detection a matter of concern. Ghalechyan et al. (2024) developed probabilistic neural networks to offer confidence levels in addition to the 97% level of accuracy needed in production settings. Bozkir et al. (2022) enhanced embedding methods through GramBeddings to leverage n-grams in the quadruplet neural networks to offer an accuracy level of 98.27% that is language-insensitive.

2.2 Transformer-Based and Large Language Model Detection

Transformer architectures transformed phishing detection by modeling long-range dependencies and contextual relationships within URL strings. Wang et al. (2023) presented PhishBERT, a specialized variant of BERT pre-trained on three billion URLs utilizing manipulated token detection objectives, attaining exceptional performance with over 99.98% accuracy. This pre-training task was domain-specific and proves the importance of domain-specific adaptability in the transformer model in security applications. The URLTran model was designed and developed by Maneriker et al. (2021) study, demonstrating a true positive rate of 86.80% and a false positive rate of 0.01%. This was a 21.9% improvement over previous baselines at this operating threshold. The significant performance disparity between PhishBERT's specialized training and URLTran's general methodology underscores the trade-offs between pre-training investment and deployment adaptability.

Current studies have analyzed the application of large language models in phishing identification, and the results show the strengths and weaknesses in the current applications of large language models in phishing identification. For instance, a study conducted by Patel et al. 2024 examined the application of fifteen large language models in phishing identification, and the best models with scores ranging from 8 to 10 include ChatGPT 3.5, GPT-3.5-Turbo-

Instruct, and ChatGPT 4. Mahendru and Pandit (2024) systematically compared the results and found DeBERTa V3 outperformed GPT-4 with a significantly higher recall of 95.17% against GPT-4's 91.04% in the same evaluation test, thereby implying that large language models exhibit a higher level of semantic understanding but might perform worse in specific objectives when fine-tuned models are utilised. Ye et al. (2023) described the GPT-3.5 capability profile, noting its excellent semantic understanding along with variability in its performance in structured pattern identification tasks, thus supporting the notion that the semantic capabilities of GPT-3.5 and the structural capacity of the BERT model complement each other.

The crucial restriction in the transformer-based approaches is their need for structural URL assessment in the absence of any world knowledge required for the identification of sophisticated attacks concerning the impersonation of brand names and their URLs. The two approaches, PhishBERT and URLTran, fail to recognize any trustworthy domains having a similar sounding name to the renowned organization because structural assessment will be inadequate if the attacker has a persuasive domain name with a similar structure to the trustworthy domains.

2.3 Hybrid Architectures and Research Gap

The fusion of several architectures to tap the strengths of each has shown encouraging results in other fields. Dhiman et al. (2024) introduced the GBERT model, a combination of the GPT-BERT architectures designed for fake news identification with an accuracy of 95.30%, precision of 95.13%, and a recall value of 97.35%. The GBERT model combined the embeddings of the two models by concatenating them and proved that the combination of heterogeneous embeddings outperforms the individual models. Nevertheless, the concatenation method considers the two embedding sources in an identical manner without making an adaptive contribution adjustment, possibly hindering the optimal balance depending upon the input attributes. The efficacy of quality text embeddings derived from large language models has been substantiated to be state-of-the-art in BEIR and MTEB benchmarks by Wang et al. (2024), confirming the usability of GPT-3.5 embeddings in semantic form and proposing the efficiency enhancement of pre-trained embedding solutions.

The exhaustive literature review uncovers a substantial research deficiency at the convergence of large language model semantic comprehension and transformer-based structural analysis for phishing URL identification. PhishBERT excels in performance via domain-specific pre-training; however, it lacks the ability to utilize global knowledge for advanced brand impersonation detection. Evaluations of large language models show that they are good at understanding meaning but not as good at recognizing structural patterns as specialized models. The GBERT architecture demonstrates the viability of GPT-BERT fusion, yet employs concatenation-based integration for disinformation rather than cybersecurity applications.

The above shortcoming will be filled in the proposed study through the following three innovative aspects. First, the study will be the first to utilize the GPT-BERT model with a hybrid structure in the context of phishing URL detection. Secondly, the work will be the first to break the limitation of the concatenation methods and will introduce cross-attention mechanisms to facilitate the interaction of the embedding spaces in both directions. Finally, the work will be the first to utilize the adaptive weights in order to adjust the contribution of each embedding space according to the characteristics of the input URLs. The theory is grounded in the complementary role of GPT-3.5 embeddings in learning semantic know-how to detect contextual anomalies and social engineering attacks and the domain-specific structural

capability of the embeddings of the BERT model in learning syntactic irregularities and character-level details. The cross Attention mechanism-based fusion makes it possible for the two functions to train together and might leverage the strengths of the other to detect any possible threat that the source of the embeddings is unable to do.

3 Research Methodology

Below is the detailed description of the methodology framework that has been adopted in the design and testing phase of the GPT-3.5 and BERT embedding fusion model for phishing URLs. The methodology involves a quantitative experiment to test the efficiency of the model fusion design in comparison to the existing models.

3.1 Research Design

The study employs an empirical and experimental approach based on the assumption that the combination arrived at through the semantic embeddings of GPT-3.5 and the structural embeddings of BERT through cross Attention mechanisms is better than the phishing threat detection capability of solo embeddings. The assumption made in the present study stands true because the combination of the two architectures contains the contrasts in the attributes of the URL, and the combination of the two is a better way to assess the threat than the solo methods because the two complement each other.

The three most significant phases in the experimental environment include comparison evaluation, the fusion model training phase, and the embedding extraction phase. The embedding extraction phase includes the utilization of the GPT-3.5 and BERT models to generate two embeddings for each URL. The fusion model training phase is charged with the responsibility of developing the cross-attention pattern that will be applied in the fusion of embeddings that come from sources of a different nature. The evaluation phase determines the performance levels in terms of detection through classification metrics, and the performance levels are then compared with existing methods recorded in the literature (Safi & Singh, 2023).

3.2 Dataset Description

The study employs thePhiUSIIL Phishing URL Dataset, a diverse and large dataset sourced from the Mendeley Data Repository. The dataset was selected due to its large size, class distribution, and feature description capacity to analyze the URL well enough. Table 1 summarizes the properties of the dataset employed in the study.

Table 1: Dataset Characteristics

Attribute	Value
Dataset Name	PhiUSIIL Phishing URL Dataset
Total Samples	235,795
Benign URLs	100,945 (42.81%)
Phishing URLs	134,850 (57.19%)
Total Features	56
Distinct Domains	220,086

Distinct TLDs	695

The class distribution in the dataset is quite imbalanced, with a predominance of phishing URLs at 57.19%. Phishing attacks can potentially outweigh the number of legal URL submissions to the phishing detection platforms, and the class distribution depicts the same. Though the proposed approach relies heavily upon the URL text for embedding pattern generation and is less reliant upon the precomputed values, the feature set of 56 includes URL structure, domain, content indicators, and security indicators.

3.3 Data Partitioning Strategy

Stratified random sampling was used to divide the dataset into training, validation, and test sets. This kept the class distributions the same across all subsets. Table 2 shows a summary of how the data is split up.

Table 2: Data Split Configuration

Partition	Samples	Percentage	Purpose
Training	165,056	70%	Model parameter optimisation
Validation	35,369	15%	Hyperparameter tuning, early stopping
Test	35,370	15%	Final performance evaluation

The split ratio of 70-15-15 allows enough samples to be reserved for objective testing and validation while also ensuring the fusion model has an adequate number of samples for training and learning. The stratified sampling ensures the proportion of the first class of 42.81% benign and the second class of 57.19% phishing samples is maintained in each partition to avoid the effect of biased results due to sampling.

3.4 Embedding Extraction Methodology

The process of dual embedding extraction results in two embedding representations of each URL: semantic embeddings based on GPT-3.5 and structural embeddings based on DomURLs_BERT embeddings. This is in tandem with the latest discovery that high quality embeddings in texts collected from large language models perform well in semantic similarity learning (Wang et al., 2024).

GPT-3.5 Semantic Embeddings: The text-embedding-3-large model from OpenAI is applied in the GPT-3.5 embedding process to generate a dense vector embedding of size 3,072 dimensions. All URLs undergo an enrichment process where the raw URL text is transformed into a well-structured semantic description prior to the embedding process taking place. Through the identification of brand mentions, security indicators, and structural anomalies in a natural language form, this enhancer plays a crucial role in enabling the GPT model to leverage semantic understanding capabilities.

The Open AI embeddings API processes the descriptions of the URLs through a simple request and response process. The descriptions of the URLs are submitted to the embeddings API, and in return, a vector of size 3,072 floating-point numbers is received. For example, a URL such

as "http://paypa1-secure.com/login" would first be enriched to describe its characteristics (protocol type, brand similarity, suspicious patterns) before submission to the embedding model. The API can handle up to 2,048 inputs per request in batches, which makes it easier to work with big datasets.

DomURLs_BERT Structural Embeddings: The domain-specific model DomURLs_BERT, a variant of the BERT model that has been pre-trained on the available URL data and is accessible through the HuggingFace model archive, is employed in the embedding extraction process in the BERT model. The 768-dimensional embeddings generated by the model capture the character-level relations and URL structure present in URL strings. The model has the capability to directly capitalize on the URL structure because the embeddings are directly derived from URLs without any supplementation, unlike the GPT methods.

The extraction process involves the tokenization of each URL through the application of the BERT tokenizer with a maximum sequence length involving 512 tokens, followed by the application of the encoder to produce the contextual embeddings, and the collection of the embeddings of the [CLS] token representing the aggregated URL embedding. The process adopts the commonly followed best practices in the application of the BERT model in the classification process (Wang et al., 2023).

3.5 Experimental Environment

The tests were done on Google Colab Pro, which had an NVIDIA A100 GPU that sped things up and had 40 GB of GPU memory. PyTorch 2.0 is used as the deep learning framework, HuggingFace Transformers 4.x to get to the BERT model, and the OpenAI API to make GPT-3.5 embeddings. The DomURLs_BERT model is retrieved from the HuggingFace model hub (amahdaouy/DomURLs_BERT). This setup gives GPU-accelerated computation, which is necessary for training the neural fusion architecture and working with the large dataset.

3.6 Evaluation Metrics

The evaluation framework uses standard classification metrics that are good for binary phishing detection tasks. Table 3 shows the metrics that were used to measure performance.

Table 3: Evaluation Metrics

Metric	Definition
Accuracy	(TP + TN) / (TP + TN + FP + FN)
Precision	TP / (TP + FP)
Recall	TP / (TP + FN)
F1-Score	2 × (Precision × Recall) / (Precision + Recall)
AUC-ROC	Area under Receiver Operating Characteristic curve
False Positive Rate	FP / (FP + TN)

All the above measures provide a combined evaluation of the detection capacity and reliability of the system through the detection capability and reliability, measured by the F1-score. The rate of false positive results is a concern because of its significance in the operational

environment, where high rates of false alerts will undermine the usability and credibility of the system (Gupta et al., 2021).

4 Design Specification

This section will introduce the architectural design of the proposed GPT-3.5 and BERT embedding fusion system, describing the elements and methods involved in realizing the effective fusion of GPT-3.5 semantic embeddings and BERT structural embeddings for phishing URL identification.

4.1 System Architecture Overview

The proposed solution involves a three-stage pipeline design in terms of embedding extraction, cross-attention fusion, and classification tasks. The entire embedding preparation pipeline is illustrated in Figure 1, along with the parallel process involved in the generation of GPT-3.5 and the BERT embeddings in the construction process.

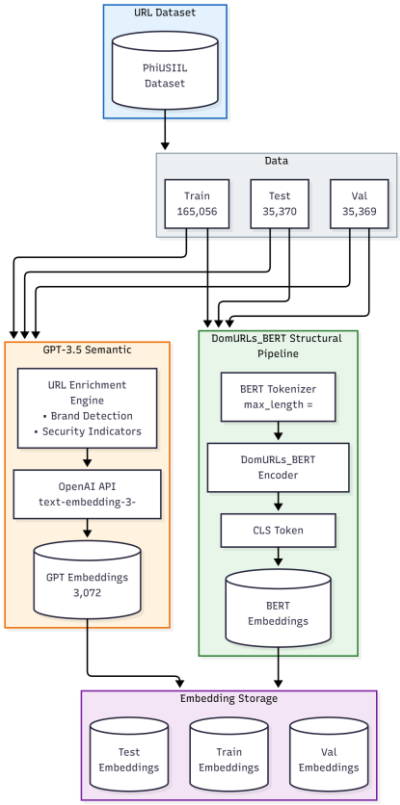

Figure 1: Dual-Embedding Preparation Pipeline illustrating the extraction of GPT-3.5 semantic embeddings (3,072 dimensions) through URL enrichment and OpenAI API, alongside DomURLs_BERT structural embeddings (768 dimensions) through tokenization and encoder processing. The pipeline processes the PhiUSIIL dataset (235,795 URLs) split into training (165,056), validation (35,369), and test (35,370) partitions, with final embeddings stored for fusion model training.

Figure 1 shows that the architecture involves the separation of embedding extraction and the process of training the fusion model to make way for the precomputation and caching of embeddings to minimize computational costs involved in the training process of the model. The two-branch architecture involves the processing of the URL through the GPT-3.5 semantic pipeline and the DomURLs_BERT structural pipeline, represented by the left branch with an orange border and the right branch with a green border, respectively.

4.2 GPT-3.5 Semantic Embedding Branch

The GPT-3.5 branch uses a two-step process to turn raw URLs into semantically rich embeddings. The first step is URL enrichment, and the second step is API-based embedding generation, as shown in the left part of Figure 1.

URL Enrichment Engine: The raw URL strings are converted to a natural language description to convey the security attributes through the enrichment process. This enhances the capability of the GPT model to utilize its semantic understanding capability to analyze the URL, where the raw URL strings provide less context to the URL analysis process, as discussed in Table 4.

Table 4: URL Enrichment Components

Component	Analysis Performed	Output
Protocol Analysis	HTTP vs HTTPS detection	Security status indicator
Brand Detection	Financial and technology brand identification	Brand risk assessment
Domain Analysis	Length, character composition, IP address presence	Structural anomaly flags
Security Keywords	Login, verify, secure, update pattern detection	Social engineering indicators

The URL enrichment engine scans each URL for phishing pattern characteristics, such as the presence of the word (references to PayPal, Microsoft, Google, and similar high-value targets), security terms in a social engineering attack, and “sales and phishing characteristic measurements”, as in the “length and character distributions in URLs and query stringss.” The enhanced description contains detailed information to facilitate the GPT-3.5 model to produce distinctive embeddings.

OpenAI API Integration: The text-embedding-3-large endpoint of OpenAI processes the enhanced URL descriptions to give an embedding space of 3,072 dimensions. The API interface utilizes a batch processing method to provide maximum processing speeds while incurring the least possible cost in making API calls. The rate limiting methods ensure processing speeds are maintained while meeting the constraints for the APIs.

4.3 DomURLs_BERT Structural Embedding Branch

The BERT branch processes the URLs in their raw form with a domain-specific transformer encoder to make direct inferences based upon the syntax pattern in URLs. The DomURLs_BERT tokenizer parses each URL into subword pieces based upon a URL-content-specific vocabulary with a maximum sequence length of 512 to process long URLs efficiently. The processed sequences are submitted to the process of the 12-layer DomURLs_BERT encoder, in which self-

attention mechanisms are applied to find the structural relationships in the URL components based on the URL structure and URL components. The URL structure embedding finally utilizes the embedding of the [CLS] token to generate a 768-dimensional vector based on the attention mechanisms followed in the DomURLs_BERT encoder to capture specific structural patterns in the URLs.

4.4 Cross-Attention Fusion Architecture

The fusion architecture combines the heterogeneous embeddings of GPT-3.5 (dimension size: 3,072) and BERT (dimension size: 768) with a cross-attention mechanism that allows two-way interaction. Figure 2 shows the full architecture of the fusion model.

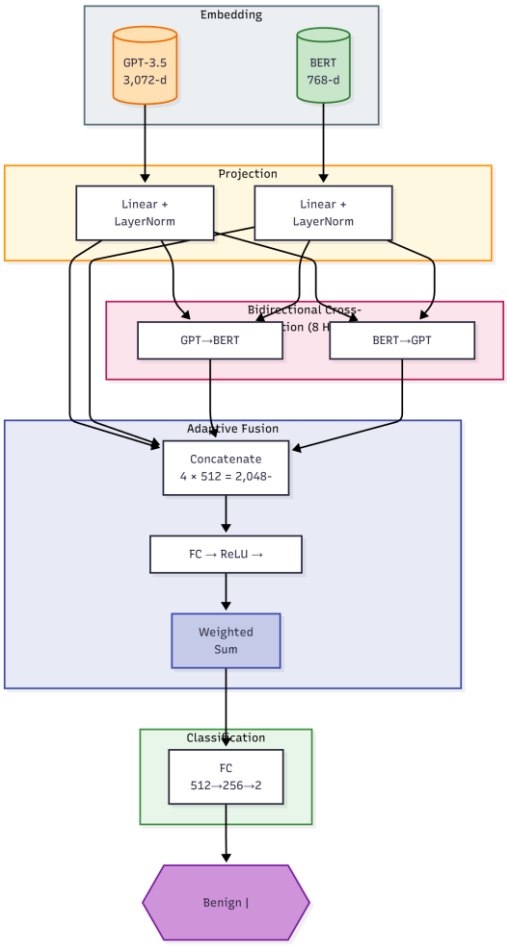

Figure 2: Cross-Attention Fusion Model Architecture showing the projection layers for dimension alignment (3,072→512 for GPT, 768→512 for BERT), bidirectional cross-attention module with 8 attention heads enabling GPT→BERT and BERT→GPT information exchange, adaptive fusion gate computing learnable weights through softmax normalisation, and classification head producing binary phishing/benign predictions.

Projection Layer: The projection layer allows for additional calculations regarding attention by aligning the dimensions of the embeddings into a unified space of size 512, shown in the top part of Figure 2. For stability in the learning process, a particular linear transformation is applied to each source embedding, followed by a layer normalization process. Apart from its regularization function, the dimensional constraint in the process forces the model to retain the most discriminative information in each source embedding.

Bidirectional Cross-Attention: The cross-attention module, denoted in Figure 2 with the pink border, consists of two different attention mechanisms: GPT-to-BERT attention, where the GPT embeddings query the BERT embeddings, and the BERT-to-GPT attention module, where the BERT embeddings query the GPT embeddings. The two attention mechanisms make use of 8 attention heads with a query, key, and value size of 512 dimensions, based on the multi-head attention equation put forward in the work of Vaswani et al., 2017.

This bidirectional architecture allows embedding sources to focus selectively on important information in the other source. The GPT-to-BERT attention mechanism allows semantic insights to be guided by structural pattern information, while the BERT-to-GPT attention mechanism allows structural analysis to consider semantic information.

Adaptive Fusion Gate: The blue-b ordered region in Figure 2 represents the fusion gate where the weights to be learned are applied to determine the contribution to the fused embedding represented by each component in the representation. The components include the GPT-projected embedding, the BERT-projected embedding, the GPT-attended-by-BERT embedding, and the BERT-attended-by-GPT embedding components. The four embeddings are each 512-dimensional and are concatenated to produce a 2,048-dimensional input to the fusion gate network.

The four output values are generated through the gating network in a fully connected layer followed by ReLU activation and a final projection layer. The softmax normalization term guarantees the weights assigned to the contribution percentages are understandable by normalizing the values to weights summing to one. The fused embedding of size 512 dimensions is formed based on the weights of the four components of the representation weights.

4.5 Classification Head

The classification head uses a three-layer fully connected network to produce binary class predictions from the fused embedding, and this is illustrated in the green bordered box at the bottom of Figure 2. The first layer has a projection from 512 to 512 dimensions with GELU activation and dropout (probability 0.3). The second layer reduces the dimension to 256 with the same activation function and regularization. The output layer projects to a 2-dimensional space corresponding to the classes benign and phishing logits with softmax activation for probability calculation. The GELU activation function is known to provide smooth non-linearity and has been effective in transformer networks (Hendrycks & Gimpel, 2016). Layer normalization is applied to each dense layer to improve the flow of the gradient and enhance convergence.

5 Implementation

This part talks about how the proposed hybrid embedding fusion system was technically put together, with a focus on the training setup and methods used to get the reported performance metrics.

5.1 Fusion Model Architecture and Training

The fusion model is realized through the use of the PyTorch 2.0 architecture based on custom modules of cross-attention and the adaptive fusion gate. The implementation has in excess of 2.3M trainable parameters, which are implemented through the projection layers, attention mechanisms, fusion gates, and classification heads. The model adopted the dropout regularisation technique with a probability of 0.3 to avoid overfitting.

The model was trained with the AdamW optimiser, learning rate of 1×10^{-4}, and weight decay of 1×10^{-5}. The batch size was set to 256 while training for a maximum of 30 epochs, making use of early stopping with patience of 7 epochs based on the F1-score value on the validation data. The OneCycleLR learning rate scheduler consisted of a learning rate schedule with warm-up phase followed by a rise to the maximum learning rate and finally an anneal to the minimum value, making it ideal for having aggressive learning rates in the middle phase for the convergence of transformer-based models (Smith & Topin, 2019).

The cross-entropy loss function with class weights inversely proportional to class occurrences was the training criterion, treating the moderate class imbalance in the dataset (57.19% phishing and 42.81% benign) in a compensatory manner. The choice of weights eliminates model class bias against the phishing class while ensuring the correctness of benign URL detection. The random seeds in stochastic processes are fixed for comparability across architectural changes.

5.2 Training Dynamics and Convergence

The training involved the calculation of the forward pass-through embedding projections, cross-attention mechanisms, fusion gates, and classification heads while iterating through the training samples of size 165,056 in mini-batch chunks. Roughly, each epoch involved approximately 645 steps, and model evaluation through the calculation of performance metrics occurred at the end of each epoch.

The training convergence took place in the first 15 epochs, while the best F1-score in the validating dataset was reached in the 12th epoch. The early stopping function triggered activation in the 19th epoch because of a stagnant period of improvement for seven consecutive epochs, and the model weights were changed to those of the 12th epoch to make the final prediction assessment.

The embedding extraction pipeline applied the entire dataset to the precomputation, resulting in embedding matrices of shapes (165,056×3,072) and (165,056×768) for GPT-3.5 and BERT respectively in the training split. The corresponding embedding matrices for the validation and testing splits had similar proportional dimensions. The precomputation and embedding serialization reduced the necessity to preserve the entire transformer model in the GPU memory while training the fusion model.

The assessment of the quality of embedding quality before the fusion model training proved the capture of the discriminative information in both sources. The superiority of the embeddings of DomURLs_BERT in class separability with the score value of 0.1017 and the separability ratio of 0.4189 against the GPT-3.5 embeddings with the score value of 0.0323

and the separability ratio of 0.2286 respectively supported the capture of the discriminative information in both sources in favor of the fusion assumption.

6 Evaluation

This section offers a detailed discussion of the results of the experiment conducted by the hybrid GPT-3.5 and BERT embeddings fusion system for the purpose of phishing URL identification. The evaluation evaluates the performance of the model in a variety of dimensions, and the understanding and implications of the results are facilitated by the implementation of statistics and visual illustrations.

6.1 Experiment 1: Overall Model Performance

The first experiment uses a test set of 35,370 URLs that were not used in the training process to see how well the fusion model can classify them. This assessment provides the primary evaluation of the proposed architecture's effectiveness and addresses the third research objective concerning detection performance benchmarks.

The fusion model was run on the test samples with frozen parameters, and the predictions are evaluated against the ground truth labels to calculate the classification metrics. Figure 3 shows the classification metrics for the proposed fusion model.

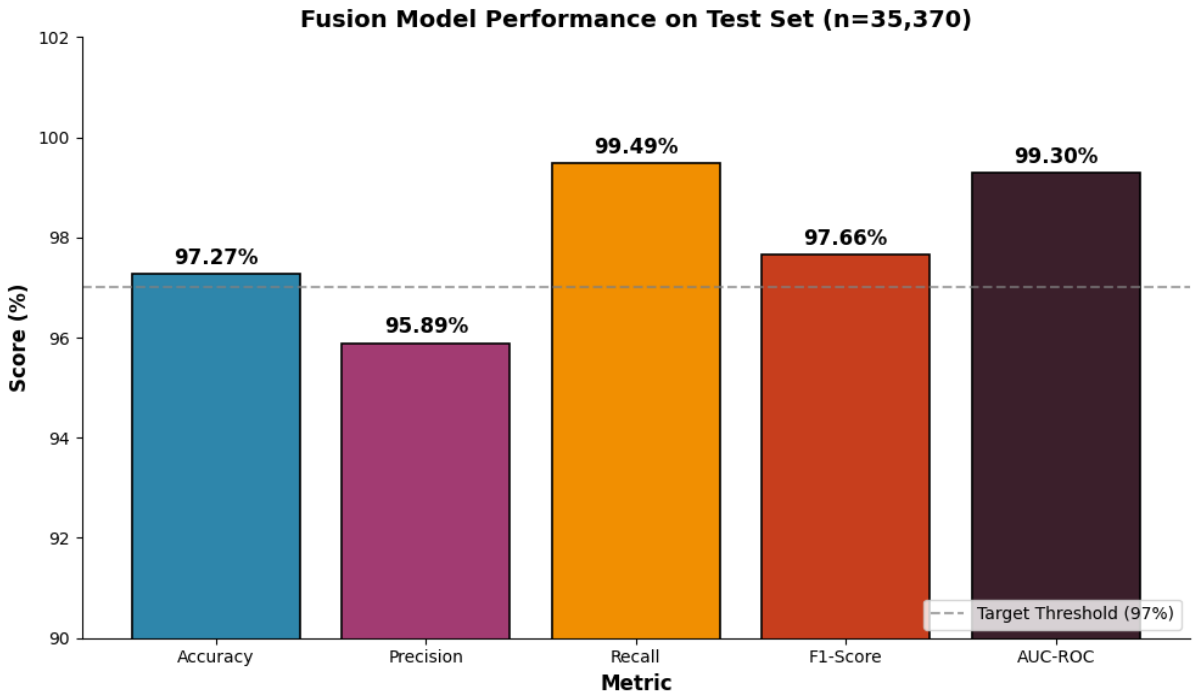

Figure 3: Performance metrics of the fused model on the test data (n=35,370). The model has an accuracy of 97.27%, precision of 95.89%, recall of 99.49%, F1 score of 97.66%, and AUC ROC of 99.30%;

The results indicate the effectiveness of the proposed fusion model in the classification process based on each evaluated criterion. The model with an accuracy rate of 97.27% satisfies the target threshold of 97% in the objectives of the study to correctly categorize many URLs in the test data. The high recall value of 99.49% in the results is a crucial point in the phishing detection tool, where only 0.51% of the phishing URLs are able to pass undetected. The value of precision at 95.89% shows the rate of the predicted phishing URLs that are indeed phishing URLs, where the remaining 4% of the positive results are a false positive rate in the positive class. The F1-score value of 97.66% shows a well-rounded performance in terms of both Precision and Recall rates, substantiating the efficiency of the model. The value of AUC-ROC at 99.30% assures superior discrimination skills in terms of the classification thresholds.

6.2 Experiment 2: Confusion Matrix and Error Analysis

The second experiment offers a thorough understanding of the kinds of errors generated by the fusion model through the examination of the results of classification performed through confusion matrix analysis. The need to understand the real-world implications of model application and the possible fields for improvement stems from this requirement. Figure 4 presents the confusion matrix visualisation with absolute counts and percentage annotations.

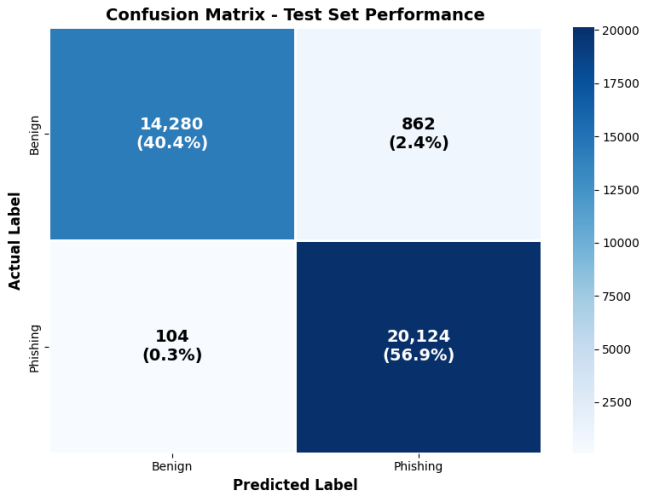

Figure 4: Confusion matrix for the fusion model on the test set. The model correctly identifies 14,280 benign URLs (true negatives) and 20,124 phishing URLs (true positives), whilst producing 862 false positives and 104 false negatives.

There is a mismatch in the number of occurrences of the two categories in the classification results based on the results from the confusion matrix analysis. The value of the false positive rate, which stands at 5.69%, is a result of the model producing a total of 862 samples in the false positive category, known as harmless URLs incorrectly labeled as phishing URLs. The value of the false negative rate, standing at 0.51%, is a result of the model producing a total of 104 samples in the false negative category, referred to as phishing URLs that go undetected.

The number of true phishing URLs detected is 20,124 out of a total of 20,228 phishing URLs, establishing the effectiveness of the model in identifying and pointing out phishing content with a true positive rate of 99.49%. The number of true negative URLs is 14,280 out of a total of 15,142 benign URLs. Practically, a low rate of false negatives guarantees the user a high degree of protection against phishing attacks, while the rate of false positives of 5.69% will need to be fine-tuned based upon the application requirements if direct verification is required based upon the results.

6.3 Experiment 3: Fusion Weight Distribution Analysis

To shed light upon how the model weighs the contribution of the different embedding sources, the third experiment investigates the learned fusion weights. This investigation reveals the importance of semantic versus structural embeddings and corresponds to the answer to the fourth research question in terms of proving the advantage of fusion.

Figure 5 presents the distribution of fusion weights across the four representation components.

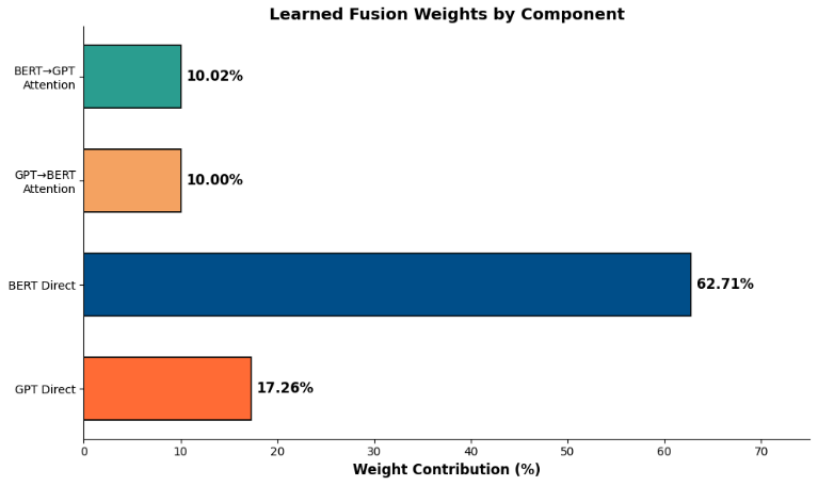

Figure 5: Distribution of learned fusion weights across the four representation components. BERT direct embeddings dominate with 62.71% contribution, whilst GPT direct embeddings contribute 17.26%. Cross-attention components (GPT→BERT and BERT→GPT) each contribute approximately 10%.

The fusion weight analysis shows that the model gives much higher weights to the direct embeddings of the BERT model (62.71%) than the GPT direct embeddings (17.26%). The dominance in structural representation is in accordance with the class separability values given in Section 5.2, where the DomURLs_BERT embeddings outperformed the GPT-3.5 embeddings in terms of class separability values. The domain-specific pre-training of the DomURLs_BERT embeddings with URL datasets shows a higher ability to be discriminative in the phishing domain than the GPT embeddings.

Despite the dominance of the BERT results, the contribution of GPT at 17.26% shows that semantic embeddings offer complementary information that the model is able to leverage for the classification task. The contribution of the cross-attention components is approximately

10% (GPT→BERT: 10.00%, and BERT→GPT: 10.02%), and this shows that the bidirectional attention flow allows for the interaction of information from the embedding sources. The total contribution of the GPT components related to the architecture (direct and the two cross-attention components) comes to 37.28%, and this shows that the architecture is able to leverage the two sources of embedding instead of settling for a solution from a single model.

6.4 Discussion

The experimental findings confirm the research hypothesis that the combination of GPT-3.5 semantic embeddings and BERT structural embeddings through cross-attention layers improves phishing URL identification. The results show a promising level of accuracy at 97.27%, F1-score of 97.66%, and an AUC-ROC value of 99.30%. The excellent recall rate of 99.49% is indeed important in a security context, where almost every phishing attempt should be caught while keeping the levels of precision reasonable.

Fusion weight analysis shows the contribution of each embedding source to the hybrid model architecture. The contribution of the embeddings shows that the BERT embeddings make the highest contribution with a value of 62.71%, aligning with the success of the PhishBERT model (Wang et al., 2023), due to the domain-specific pretraining of the model on URL data. The GPT embeddings contribute of 17.26%, thus aligning with the fact that semantic understanding complements structural analysis. The cross-attention components add up a total of 20% (GPT→BERT: 10.00%, BERT→GPT: 10.02%), thus confirming the two-way flow of information from the embedding spaces. The dynamic weighing system outshines the concatenation-based fusion systems, for example, GBERT (Dhiman et al., 2024), in the way it balances structural and semantic information depending upon the

Comparative assessment against the existing literature yields a competitive but differentiated performance profile. The baseline results based on a single model indicate a higher level of accuracy: the Random Forest approach based on the word-level feature achieved a value of 99.57% (Gupta et al., 2021), the 1D CNN model reached a value of 99.70% (Haq et al., 2024), while the PhishBERT model recorded a value of 99.98% through domain-specific pre-training (Wang et al., 2023). The merit of the current model's results in terms of its accuracy is relatively marginal at 97.27. The architectural contribution resides within the achievement of the GPT-BERT combination in the cyber security domain and the implementation of the cross-attention mechanisms to dynamically weight the contribution instead of the concatenative model.

There are a number of shortcomings that need to be noted in an effort to improve in the future. The false positive rate value of 5.69% is above the production levels, and this might call for the need to calibrate the production thresholds or implement a two-step verification process. The high FPR is a characteristic that depicts the trade-off made in favor of sensitivity and to the detriment of specificity. The URL enrichment process only accounted for the training data and might limit the discriminative feature of GPT embeddings in the remaining partitions regarding semantic contribution assessment. The embedding quality assessment proved the weakness of class separability in GPT-3.5-based embeddings against the quality of BERT embeddings, denoted by the respective silhouette values of 0.0323 and 0.1017. The results should be verified and generalized through testing the model on various phishing URL datasets and in the context of its application to a zero-day attack scenario.

7 Conclusion and Future Work

This study examined the integration of GPT-3.5 semantic embeddings with BERT structural embeddings via a cross-attention mechanism to improve phishing URL detection. The first research question investigated was: *How does the fusion of GPT-3.5 semantic embeddings with BERT contextual embeddings through cross-attention mechanisms enhance phishing URL detection accuracy compared to individual embedding approaches?* This question arose from a recognized deficiency in the literature, where current methodologies predominantly utilize single-model architectures, inadequately capturing both semantic intent and structural patterns in malicious URLs.

7.1 Achievement of Research Objectives

Four objectives in the research study were formed to tackle the question in a detailed manner. The first objective involved the creation of a two-step embedding extraction pipeline to produce GPT3.5 semantic embeddings and structural embeddings through the URL input in a way that the pipeline utilizes the OpenAI text-embedding-3-large API URL embedding technique to produce embeddings of size 3,072 dimensions and the domain-specific model DomURLs_BERT to produce embeddings of size 768 dimensions through the processing of 235,795 URLs combined across the partitions. The second aim was to design and implement a cross-attention fusion architecture that made it possible to exchange information in both directions from the embedding spaces to be fused. The special design comprising projection layers for aligning the dimensions, an 8-head bidirectional cross-attention module, an adaptive fusion gate with softmax weights to be learned, and a classification head to generate understandable weight distributions in fusion was employed to accomplish this second aim.

The third aim in this case was the achievement of a detection performance with a rate of accuracy above 97%. The test results validate the accomplishment of this aim, where the achieved accuracy, F1-score, AUC-ROC, and most importantly the recall in the test dataset of 35,370 URLs stood at 97.27%, 97.66%, 99.30%, and 99.49%, respectively.

The fourth aim was to demonstrate the complementary strengths of multi-modal fusion through an analysis of contribution weights learning results. The contribution percentages of the direct embeddings of BERT, the direct embeddings of GPT, and the cross-attention components are 62.71%, 17.26%, and around 20% in total, respectively, based on the fusion weights, proving that both sources indeed make a contribution to the classification and do not fall into a single-source solution.

7.2 Key Findings and Implications

The work yields a number of important results that improve the understanding of the process of embedding fusion in the area of cybersecurity problems. The key result proves the concept of the fusion of model architectures to solve the URL classification task because the cross-attention fusion of GPT-3.5 and BERT embeddings attains competitive phishing detection results. Domain-specific embeddings in the BERT model contribute the strongest discriminative information according to the fusion weights, while the general GPT models provide the strongest semantic complement. The results indicate that the strongest combined

models will involve domain-adapted modules along with the integrated language model modules in the combined architectures.

Academically speaking, the work applies architectural concepts from a related domain to the area of cyber security and demonstrates the first practical application of GPT-BERT combined architecture designed specifically for phishing URL detection tasks in particular. Moving past the traditional concatenation methods, the cross-attention fusion approach offers a practical contribution to methodology that allows for the dynamic weighing of the differing embedding sources. Practically speaking, the interpretable weights in the fusion function make it easier to implement in security applications, and the corresponding high value of 99.49% implies a robust protective capability for users.

7.3 Limitations

The output is affected by a variety of issues. Only the training data was evaluated in the URL enrichment process, and this may tend to undervalue the importance of GPT embeddings. The fact that the rate of false positives of 5.69% is higher than typical values mean the need for calibration in terms of the production environment. The assumption regarding generalizability and processing costs of two embeddings per sample to provide scaling for high-throughput applications is a limitation in terms of the evaluation environment.

7.4 Future Work

There are a number of important research directions that can be derived from the above study. The first and most important one is the assessment based on the zero-day attack instances via the temporal split of the dataset, where the model will be trained based on the historical collection of URLs and the testing phase based on the phishing instances that will be developed in the future.

The development of live browser extension implementation to evaluate its effectiveness in real phishing attacks constitutes another area of study. Gaining insight into the practical efficacy and acceptability of the system in the real environment, the results of which would be difficult to determine in a lab test, may be possible through a controlled trial in conjunction with cyber security professionals. The implementation process may also generate learning data that will be continually updated based on attack trends.

The study of end-to-end fine-tuning is another promising area of investigation. The current work uses frozen pre-trained embeddings, and fine-tuning both embedding models for the phishing detection task may improve discriminatory power and rebalance the weights of fusion towards a more equitable contribution of semantic and structural views.
Finally, a complete threat assessment could be accomplished through the expansion of multi-modal threat detection to encompass similarity assessments concerning rendered phishing pages, certificate transparency attributes, and WHOIS registration trends. The development environment for the complete phishing detection solution has its basis in the adaptive fusion gate architecture, thus facilitating the incorporation of any additional input modes.

7.5 Concluding Remarks

The current study has successfully demonstrated the effectiveness of a phishing URL detection model designed by combining GPT-3.5 semantic embeddings and BERT structural embeddings through the mechanism of cross-attention in the task of phishing URL detection because the proposed architecture has achieved an accuracy of 97.27% and a Recall of 99.49% and has met the entire learning task objectives defined in the set study.

Speaking in the context of the research question, the data reveals that the combination of GPT-3.5 semantic embeddings and BERT contextual embeddings through the application of cross-attention mechanisms indeed has a positive effect upon the identification of phishing URLs by allowing the model to leverage the complementarities in the two embeddings to capture structural and semantic clues respectively. The application of cross-attention allows the learned weights to ensure the significance of both sources. By offering a renewed architectural perspective and scientific proof for the hybrid embedding methods in security-critical classification problems, the proper utilization of GPT-BERT Fusion in the cyber security domain has prepared the way for the forthcoming multi-modal fusion strategies in threat identification applications.

References

Bozkir, A. S., Dalgic, F. C., & Aydos, M. (2022). GramBeddings: A new neural network for URL based identification of phishing web pages through N-Gram embeddings. *Computers & Security*, 124, 102964. https://doi.org/10.1016/j.cose.2022.102964

Da Silva, C. M. R., Feitosa, E. L., & Garcia, V. C. (2019). Heuristic-based strategy for phishing prediction: A survey of URL-based approach. *Computers & Security*, 88, 101613. https://doi.org/10.1016/j.cose.2019.101613

Dhiman, P., Kaur, A., Gupta, D., Juneja, S., Nauman, A., & Muhammad, G. (2024). GBERT: A hybrid deep learning model based on GPT-BERT for fake news detection. *Heliyon*, 10(16), e35865. https://doi.org/10.1016/j.heliyon.2024.e35865

Ghalechyan, H., Israyelyan, E., Arakelyan, A., Hovhannisyan, G., & Davtyan, A. (2024). Phishing URL detection with neural networks: An empirical study. *Scientific Reports*, 14(1). https://doi.org/10.1038/s41598-024-74725-6

Gupta, B. B., Yadav, K., Razzak, I., Psannis, K., Castiglione, A., & Chang, X. (2021). A novel approach for phishing URLs detection using lexical based machine learning in a real-time environment. *Computer Communications*, 175, 47–57. https://doi.org/10.1016/j.comcom.2021.04.023

Hannousse, A., & Yahiouche, S. (2021). Towards benchmark datasets for machine learning based website phishing detection: An experimental study. *Engineering Applications of Artificial Intelligence*, 104, 104347. https://doi.org/10.1016/j.engappai.2021.104347

Haq, Q. E. U., Faheem, M. H., & Ahmad, I. (2024). Detecting phishing URLs based on a deep learning approach to prevent cyber-attacks. *Applied Sciences*, 14(22), 10086. https://doi.org/10.3390/app142210086

Hendrycks, D., & Gimpel, K. (2016). Gaussian error linear units (GELUs). *arXiv*. https://arxiv.org/abs/1606.08415

Mahendru, S., & Pandit, T. (2024). SecureNet: A comparative study of DeBERTa and large language models for phishing detection. *arXiv*. https://doi.org/10.48550/arxiv.2406.06663

Maneriker, P., Stokes, J. W., Lazo, E. G., Carutasu, D., Tajaddodianfar, F., & Gururajan, A. (2021). URLTran: Improving phishing URL detection using transformers. *arXiv*. https://arxiv.org/abs/2106.05256
Nanda, M., Saraswat, M., & Sharma, P. K. (2024). Enhancing cybersecurity: A review and comparative analysis of convolutional neural network approaches for detecting URL-based phishing attacks. *e-Prime - Advances in Electrical Engineering Electronics and Energy*, 8, 100533. https://doi.org/10.1016/j.prime.2024.100533
Patel, H., Rehman, U., & Iqbal, F. (2024). Large language models spot phishing emails with surprising accuracy: A comparative analysis of performance. *arXiv*. https://arxiv.org/abs/2404.15485
Safi, A., & Singh, S. (2023). A systematic literature review on phishing website detection techniques. *Journal of King Saud University - Computer and Information Sciences*, 35(2), 590–611. https://doi.org/10.1016/j.jksuci.2023.01.004
Smith, L. N., & Topin, N. (2019). Super-convergence: Very fast training of neural networks using large learning rates. *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, 11006, 1100612. https://doi.org/10.1117/12.2520589
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998–6008.
Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R., & Wei, F. (2024). Improving text embeddings with large language models. *arXiv*. https://arxiv.org/abs/2401.00368
Wang, Y., Zhu, W., Xu, H., Qin, Z., Ren, K., & Ma, W. (2023). A large-scale pretrained deep model for phishing URL detection. *ICASSP 2022 - IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. https://doi.org/10.1109/icassp49357.2023.10095719
Ye, J., Chen, X., Xu, N., Zu, C., Shao, Z., Liu, S., Cui, Y., Zhou, Z., Gong, C., Shen, Y., Zhou, J., Chen, S., Gui, T., Zhang, Q., & Huang, X. (2023). A comprehensive capability analysis of GPT-3 and GPT-3.5 series models. *arXiv*. https://arxiv.org/abs/2303.10420