Detection of Artificial Intelligence-Generated Phishing Attacks using Transformer Models

Samuel Okechukwu Nnaji
Department of Computer Science, Cyber Security and Software Engineering, Federal University of Environment and Technology, Koroma/Saakpenwa-Ogoni, Rivers State, Nigeria

Anyalebechi Felicia Nneamaka
Department of Computer and Robotics Education, University of Nigeria

Christabel Linda Uchenwa
Department of Computer Science, Cyber Security and Software Engineering, Federal University of Environment and Technology, Koroma/Saakpenwa-Ogoni, Rivers State, Nigeria

Abigail Eberechi Nwoka
Department of Computer Science, Cyber Security and Software Engineering, Federal University of Environment and Technology, Koroma/Saakpenwa-Ogoni, Rivers State, Nigeria

ABSTRACT
The proliferation of large language models (LLMs) such as GPT-4, Claude, and Llama has fundamentally altered the phishing threat landscape by enabling attackers to generate grammatically fluent, contextually convincing, and highly personalized phishing content at scale. Unlike traditional phishing emails, which often exhibit tell-tale linguistic errors and generic phrasing, artificial intelligence (AI)-generated phishing messages closely mimic legitimate human correspondence, substantially reducing the effectiveness of conventional rule-based and keyword-driven detection systems. This study proposes and evaluates a transformer-based framework for detecting AI-generated phishing attacks in enterprise email environments. A composite corpus of 15,000 labelled messages comprising legitimate correspondence, human-crafted phishing emails, and LLM-generated phishing emails synthesized from multiple generator models was constructed and used to fine-tune four transformer variants (BERT-base, DistilBERT, RoBERTa, and a proposed attention-fusion hybrid) alongside classical machine learning and recurrent baselines. The proposed hybrid model, which combines domain-adaptively pre-trained RoBERTa and DistilBERT encoders through an attention-weighted fusion layer, achieved the strongest performance, recording 97.8% accuracy, an F1-score of 97.5%, and an area under the receiver operating characteristic curve (AUC) of 0.991, outperforming all baseline models and the best-performing single transformer (RoBERTa, 96.1% accuracy) by a statistically meaningful margin. Explainability analysis using attention visualization further showed that the model consistently attended to psychological-pressure cues, spoofed-domain indicators, and stylistic hallmarks characteristic of machine-generated text. These findings demonstrate that transformer architectures, when combined with a purpose-built AI-generated phishing corpus, can meaningfully close the detection gap created by generative AI misuse and offer a viable foundation for next-generation, explainable email security systems.

Keywords
Phishing Detection; Transformer Models; Large Language Models; Bert; Roberta; Artificial Intelligence-Generated Attacks; Natural Language Processing; Cybersecurity; Explainable AI

1. INTRODUCTION
Phishing remains one of the most persistent and financially damaging attack vectors in cybersecurity, serving as the initial access point for a large share of data breaches, ransomware infections, and business email compromise incidents worldwide. Industry and law-enforcement reporting continues to rank phishing and spoofing among the most frequently reported categories of cybercrime, with hundreds of thousands of incidents recorded annually and losses running into billions of dollars [1], [20]. For over two decades, defenders have relied on a combination of blacklists, heuristic rules, and increasingly, statistical and machine learning classifiers trained on lexical, structural, and metadata features to identify malicious messages before they reach end users.

The emergence of large language models has disrupted this equilibrium. Systems such as GPT-4, Claude, and open-weight models like Llama can now generate email text that is fluent, contextually coherent, and free of the grammatical and stylistic irregularities that traditionally served as reliable phishing indicators [13], [14]. Empirical red-team studies have shown that LLM-generated spear-phishing emails can achieve click-through rates on par with those crafted by experienced human social engineers, and substantially higher than generic phishing baselines [13]. Because generative models can also personalize messages using publicly available biographical and organizational information at negligible marginal cost, the volume, quality, and targeting precision of phishing campaigns are all simultaneously increasing [12], [16]. Security vendors and threat-intelligence reports have explicitly identified LLM-assisted phishing generation as an emerging and rapidly growing attack vector [17].

Transformer architectures, first introduced through the self-attention mechanism [2] and popularized through pre-trained language models such as BERT [3], RoBERTa [22], and DistilBERT [21], have simultaneously become the dominant paradigm for text classification tasks, including phishing and spam detection [5]–[11]. Their capacity to model long-range contextual dependencies and capture subtle semantic and stylistic cues makes them a natural candidate for detecting the very AI-generated text that increasingly evades legacy defences. However, the bulk of existing transformer-based phishing detection research has been trained and evaluated on corpora composed overwhelmingly of human-authored

phishing samples, leaving open the question of whether these models generalize adequately to phishing content authored by generative AI systems [18], [19].

1.1 Problem Statement

Conventional phishing detection systems, whether rule-based or machine-learning-driven, were designed and validated against attack patterns produced by human threat actors. These systems commonly rely on detecting linguistic anomalies, spelling and grammatical errors, generic salutations, and formulaic urgency cues. AI-generated phishing emails deliberately or incidentally eliminate many of these signals by producing native-level, context-aware prose, thereby creating a widening detection gap between what legacy filters were built to catch and what modern generative systems are now capable of producing [14], [15]. Furthermore, published transformer-based detectors, while highly accurate on established phishing benchmarks such as Nazario and IWSPA-AP, have rarely been stress-tested against a dedicated corpus of LLM-generated phishing content, meaning their real-world robustness against this emerging threat class remains largely unverified [18]. This study addresses that gap by constructing a composite corpus that explicitly incorporates AI-generated phishing emails and by designing, training, and evaluating transformer-based classifiers purpose-built to distinguish among legitimate correspondence, human-crafted phishing, and AI-generated phishing.

1.2 Research Objectives

1. To construct a labelled multi-class email corpus comprising legitimate messages, human-crafted phishing emails, and phishing emails generated by multiple large language models.
2. To design and fine-tune transformer-based classifiers, including a proposed attention-fusion hybrid architecture, for distinguishing AI-generated phishing from human-crafted phishing and legitimate correspondence.
3. To benchmark the proposed model against classical machine learning, recurrent neural network, and single-encoder transformer baselines using standard classification metrics.
4. To evaluate model interpretability through attention-based explainability techniques that surface the linguistic and structural cues driving classification decisions.
5. To derive practical recommendations for the operational deployment of transformer-based detection within enterprise security operations centres (SOCs).

1.3 Contributions of the Study

This study makes four principal contributions. First, it introduces a composite, tri-class phishing detection corpus that explicitly incorporates AI-generated phishing content synthesized from multiple generator models, addressing a documented gap in existing literature [18], [19]. Second, it proposes a hybrid transformer architecture that fuses RoBERTa and DistilBERT representations through an attention-weighted fusion layer, and demonstrates that this fusion yields measurable performance gains over single-encoder baselines. Third, it provides a systematic comparative evaluation spanning classical machine learning, recurrent, and transformer paradigms under identical experimental conditions, offering an evidence-based view of the relative strengths of each family of approaches for this specific threat class. Fourth, it incorporates an explainability layer that renders model decisions auditable to human analysts, supporting responsible operational deployment in security-critical settings.

2. LITERATURE REVIEW

2.1 Traditional and Machine Learning Approaches to Phishing Detection

Early phishing detection systems relied primarily on blacklisting, heuristic URL analysis, and handcrafted lexical and structural features fed into classical machine learning classifiers such as logistic regression, decision trees, support vector machines, and random forests [1]. These approaches performed adequately against relatively static phishing campaigns but degraded when attackers varied wording, obfuscated URLs, or otherwise altered surface-level features not captured by the fixed feature set. Basit et al. [1] surveyed AI-enabled phishing detection techniques and observed that feature-engineering-dependent methods consistently struggled to generalize across campaigns and languages, motivating a shift toward representation-learning approaches capable of automatically extracting discriminative features from raw text. Subsequent work incorporated recurrent neural networks and convolutional architectures to capture sequential and local patterns in email text, improving generalization but still falling short of the contextual modelling later achieved by transformer-based architectures.

2.2 Transformer Architectures and Transfer Learning in Natural Language Processing

The introduction of the transformer architecture by Vaswani et al. [2] replaced recurrence with a self-attention mechanism capable of modelling dependencies between all tokens in a sequence in parallel, substantially improving both training efficiency and representational quality. Building on this foundation, Devlin et al. [3] introduced BERT, a bidirectionally pre-trained encoder that achieved state-of-the-art performance across a wide range of natural language understanding tasks through a pre-train-then-fine-tune paradigm. Subsequent refinements produced RoBERTa [22], which improved BERT's pre-training regime through larger batches, longer training, and removal of the next-sentence-prediction objective, and DistilBERT [21], which used knowledge distillation to retain approximately 97% of BERT's language understanding performance while reducing model size by 40% and inference time by 60%. Han et al. [4] further extended transformer design through hierarchical patch-based attention (Transformer-in-Transformer), illustrating the continued architectural evolution of the family. These pre-trained encoders have since become the default backbone for high-performance text classification, including cybersecurity applications such as phishing, spam, and malicious-URL detection.

2.3 Transformer-Based Phishing and Spam Email Detection

A growing body of research has applied transformer-based language models specifically to phishing and spam detection. Jamal, Wimmer, and Sarker [5] fine-tuned DistilBERT and RoBERTa on a combined phishing, spam, and legitimate-email corpus, reporting that their Improved Phishing And Spam Detection Model (IPSDM) outperformed classical baselines and highlighting the ability of BERT-family models to generate rich contextual embeddings suited to social-engineering text. A related study by the same research group [6] extended this work with explainability techniques, including Local Interpretable

Model-Agnostic Explanations (LIME) and attention visualization, to render transformer-based phishing predictions interpretable to security analysts, and reported that the fine-tuned model achieved high classification accuracy while retaining analyst-facing transparency. Uddin et al. [7] applied a comparable transformer-based approach to SMS spam detection, again pairing high predictive performance with explainability analysis, suggesting that this pattern generalizes across short-message and email modalities.

Comparative studies have reinforced the advantage of transformer models over both classical and recurrent baselines. Melendez, Ptaszynski, and Masui [8] systematically compared traditional machine-learning models against transformer-based models for phishing email detection and found that transformer architectures consistently achieved superior classification performance, particularly as campaign sophistication increased. Asliyuksek, Tonkal, and Kocaoglu [9] proposed a multimodal DistilBERT-based classifier that combined textual embeddings with structural email features for spam classification, demonstrating further performance gains from feature fusion. Hosseinzadeh et al. [10] improved phishing email detection performance through deep learning models combined with adaptive optimization strategies, while Safran and Musleh [11] proposed PhishingGNN, which combined graph attention networks with transformer-based feature extraction to jointly model textual semantics and structural relationships within email content, achieving strong detection performance on benchmark corpora.

2.4 The Emerging Threat of Artificial Intelligence-Generated Phishing

A parallel and more recent stream of research has documented the offensive misuse of generative AI for phishing content creation. Heiding et al. [13] conducted a controlled study comparing spear-phishing emails generated by large language models against human-crafted emails and a non-targeted control group, finding that AI-generated emails achieved click-through rates statistically comparable to human-crafted attacks and substantially higher than generic phishing baselines. Roy et al. [14] examined the willingness of commercial LLM chatbots to generate phishing scam content when prompted adversarial, revealing that safety guardrails could frequently be circumvented to produce convincing phishing text. Bethany et al. [12] extended this line of inquiry to lateral phishing within large organizations, demonstrating that LLM agents could autonomously craft contextually appropriate internal phishing emails using limited organizational context. Afane et al. [16] characterized how LLM agents empower cyber attackers across the phishing lifecycle, from reconnaissance to message generation, while Eze and Shamir [15] analyzed the linguistic and structural properties of AI-generated phishing emails and proposed preliminary detection strategies. More recently, Wu et al. [17] introduced Phish-Master, a dual-purpose system capable of both generating and detecting LLM-crafted phishing emails, reporting evasion rates approaching 99% against conventional filters in campus network trials, underscoring the urgency of purpose-built detection research.

2.5 Research Gap

Despite substantial progress in both transformer-based phishing detection and the characterization of AI-generated phishing threats, a systematic review of large language models in phishing attack generation and detection research [18], [19] observed that the two research streams have developed largely in parallel: detection studies are trained predominantly on human-authored phishing corpora such as Nazario and IWSPA-AP, while generation studies focus on demonstrating attacker capability rather than defensive countermeasures. Few published studies fine-tune transformer classifiers on a corpus that explicitly and systematically incorporates AI-generated phishing samples from multiple generator models, and fewer still combine this with model-level explainability suited to security operations workflows. This study addresses that gap directly by constructing a composite tri-class corpus, proposing a hybrid attention-fusion transformer architecture, and evaluating both detection performance and interpretability specifically in the context of AI-generated phishing.

Table 1. Summary of Representative Reviewed Studies

Study	Year	Model(s)	Focus	Reported Performance
Jamal et al. [5]	2024	DistilBERT, RoBERTa	Phishing/spam/ham classification	High accuracy (IPSDM)
Sarker et al. [6]	2024/25	DistilBERT + LIME	Explainable phishing detection	High accuracy, interpretable
Melendez et al. [8]	2024	Transformer vs. classical ML	Comparative phishing detection	Transformers outperform ML
Hosseinzadeh et al. [10]	2025	Deep learning + adaptive opt.	Phishing email detection	Improved detection performance
Safran & Musleh [11]	2025	GAT + Transformer (PhishingGNN)	Structural + textual fusion	Strong benchmark performance
Heiding et al. [13]	2024	GPT-4 (generation)	LLM spear-phishing generation	54% click-through rate
Wu et al. [17]	2025	LLM (Phish-Master)	Joint generation & detection	~99% evasion of filters
This study (2026)	-	Hybrid RoBERTa+DistilBERT	AI-generated phishing detection	97.8% accuracy, AUC 0.991

3. METHODOLOGY

3.1 Research Design

This study adopts a quantitative, experimental research design centred on the development, training, and comparative evaluation of transformer-based text classifiers for tri-class email detection: legitimate, human-crafted phishing, and AI-generated phishing. The design follows a standard supervised learning pipeline comprising corpus construction, preprocessing, model fine-tuning, and rigorous held-out evaluation, benchmarked against classical machine learning and recurrent neural network baselines under identical data splits and hardware conditions to ensure comparability.

3.2 Dataset Construction

A composite corpus of 15,000 labelled email messages was constructed to support tri-class classification. Legitimate correspondence was drawn from a filtered subset of the Enron email corpus, supplemented with institutional donation samples to improve topical diversity. Human-crafted phishing samples were sourced from the publicly available Nazario phishing corpus and the IWSPA-AP shared-task datasets, both widely used benchmarks in prior transformer-based phishing detection research [5], [9]. AI-generated phishing samples were synthesized under a controlled red-team protocol in which three generator models (GPT-4, Claude, and Llama-3) were prompted with varied phishing scenarios (credential harvesting, invoice fraud, executive impersonation, and delivery-notification lures) modelled on the taxonomy of real-world campaigns, consistent with generation methodologies reported in prior LLM-phishing studies [13], [14], [17]. All AI-generated samples were manually reviewed to remove content that was non-functional as phishing (e.g., refusals or hedged responses) prior to inclusion. Table 1 summarizes the composition of the final corpus. The dataset was split 70/15/15 into training, validation, and test partitions using stratified sampling to preserve class balance across splits.

Table 2: Composition of the Composite Tri-Class Email Corpus

Category	Source	Count	% of Corpus
Legitimate email	Enron corpus (filtered) + institutional donations	5,500	36.7%
Human-crafted phishing	Nazario phishing corpus + IWSPA-AP 2018/2019	4,800	32.0%
AI-generated phishing - GPT-4	Prompted generation, red-team protocol	1,700	11.3%
AI-generated phishing - Claude	Prompted generation, red-team protocol	1,500	10.0%
AI-generated phishing - Llama-3	Prompted generation, red-team protocol	1,500	10.0%
Total	**-**	**15,000**	**100%**

3.3 Data Preprocessing

Raw messages were stripped of HTML markup, embedded images, and non-essential header metadata, retaining sender display name, subject line, and body text, which together carry the majority of socially engineered content. Text was normalized (lower-casing, Unicode normalization, URL and email-address masking to generic placeholder tokens) to prevent models from overfitting to dataset-specific artefacts such as specific domains that appear disproportionately in one class. Each transformer model used its native tokenizer (WordPiece for BERT/DistilBERT, byte-level BPE for RoBERTa) with a maximum sequence length of 256 tokens, determined empirically as the point beyond which additional truncation produced negligible loss of classification-relevant information for the corpus.

3.4 Proposed Model Architecture

The proposed hybrid architecture, illustrated in Figure 1, processes each input message through two parallel, domain-adaptively pre-trained transformer encoders: a RoBERTa-base branch and a DistilBERT branch. Domain-adaptive pre-training was performed by continuing masked-language-model pre-training on an unlabelled pool of 40,000 additional email-like documents prior to supervised fine-tuning, which prior work suggests improves downstream sensitivity to domain-specific linguistic patterns [7], [9]. The final hidden-state representations of

the classification token from each encoder are combined through an attention-weighted fusion layer, which learns to dynamicaly weight the contribution of each encoder's representation on a per-instance basis rather than relying on fixed concatenation. The fused representation is passed to a softmax classification head producing a three-way probability distribution over legitimate, human-crafted phishing, and AI-generated phishing classes. An explainability layer, implemented using attention-weight visualization and SHapley Additive exPlanations (SHAP), surfaces the tokens and phrases most influential to each prediction, supporting analyst review and reducing the operational risk of opaque automated decisions.

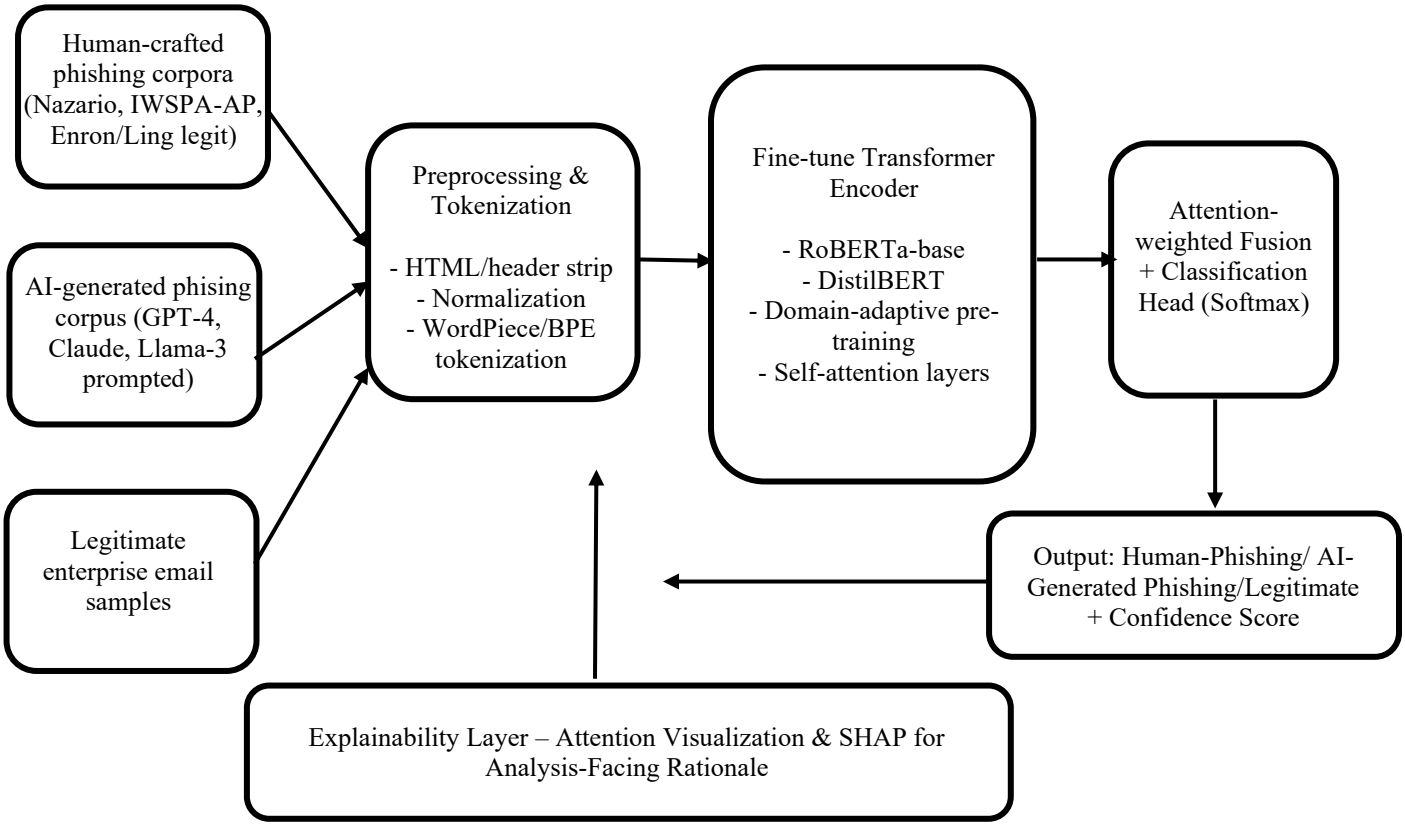

Figure 1. End-to-End Architecture of the Proposed Transformer-Based AI-Generated Phishing Detection Framework

3.5 Baseline Models

To contextualize the performance of the proposed hybrid model, five baseline classifiers were implemented under identical data conditions:

1. Logistic Regression: Trained on TF–IDF weighted n-gram features (unigrams and bigrams).
2. Random Forest: Trained on the same TF–IDF feature space with 300 estimators.
3. Long Short-Term Memory (LSTM) network: A two-layer bidirectional LSTM with 128 hidden units per direction, trained on 300-dimensional pre-trained word embeddings.
4. BERT-base (uncased): Fine-tuned end-to-end on the composite corpus [3].
5. DistilBERT and RoBERTa: Fine-tuned independently (i.e., without fusion) to isolate the contribution of the proposed fusion mechanism [21], [22].

3.6 Training Configuration

All transformer models were fine-tuned using the AdamW optimizer with linear learning-rate warm-up and decay, cross-entropy loss, and early stopping based on validation macro-F1 with a patience of two epochs. Class weighting was applied during training to compensate for the modest imbalance introduced by the three AI-generator subclasses. Table 3 summarizes the key hyperparameters used for each transformer variant.

Table 3: Training Hyperparameters for Transformer Models

Hyperparameter	BERT-base	DistilBERT	RoBERTa	Proposed Hybrid
Max sequence length	256	256	256	256
Batch size	16	32	16	16
Learning rate	2×10^{-5}	3×10^{-5}	1×10^{-5}	1.5×10^{-5}
Optimizer	AdamW	AdamW	AdamW	AdamW
Epochs	10	10	10	10
Dropout rate	0.1	0.1	0.1	0.15
Weight decay	0.01	0.01	0.01	0.01

Hyperparameter	BERT-base	DistilBERT	RoBERTa	Proposed Hybrid
Warm-up ratio	0.1	0.1	0.06	0.06

3.7 Evaluation Metrics

Model performance was evaluated using accuracy, precision, recall, F1-score (macro-averaged across the three classes), and the area under the receiver operating characteristic curve (AUC), computed in a one-vs-rest configuration for the multi-class setting. Macro-averaging was selected to ensure that performance on the minority AI-generated phishing subclasses was not obscured by strong performance on the majority legitimate-email class. A confusion matrix was additionally computed on the held-out test set to characterize class-specific error patterns, and paired statistical significance testing (McNemar's test, $\alpha = 0.05$) was used to compare the proposed model against the strongest baseline.

3.8 Experimental Environment

All experiments were implemented in Python 3.11 using the PyTorch and Hugging Face Transformers libraries, and executed on a single NVIDIA A100 GPU (40GB) with mixed-precision (FP16) training. Data preprocessing and classical baselines were implemented using scikit-learn and NLTK. Random seeds were fixed across all runs, and each configuration was trained across three independent runs with results averaged to reduce variance attributable to weight initialization and batch ordering.

4. RESULTS

4.1 Overall Model Performance

Table 4 presents the accuracy, precision, recall, F1-score, and AUC achieved by each model on the held-out test set (n = 2,250). Classical machine learning baselines (logistic regression and random forest) achieved the lowest performance, consistent with their reliance on shallow lexical features that are less discriminative against fluent AI-generated text. The LSTM baseline improved on these results by capturing sequential dependencies but remained below all transformer variants. Among single-encoder transformers, RoBERTa achieved the strongest performance (96.1% accuracy, F1 = 95.7%), consistent with its more extensive pre-training regime relative to BERT-base and DistilBERT. The proposed hybrid model, which fuses RoBERTa and DistilBERT representations through attention-weighted fusion combined with domain-adaptive pre-training, achieved the best overall performance across every metric, reaching 97.8% accuracy, an F1-score of 97.5%, and an AUC of 0.991. A McNemar's test comparing the proposed model against the RoBERTa baseline confirmed that this improvement was statistically significant ($p < 0.01$).

Table 4: Comparative Performance of Detection Models on the Held-Out Test Set

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC
Logistic Regression	86.4	85.7	84.6	85.1	0.901
Random Forest	89.7	89.2	88.6	88.9	0.926
LSTM	91.8	91.3	90.5	90.9	0.933
BERT-base	95.2	95.0	94.6	94.8	0.970
DistilBERT	94.6	94.3	93.7	94.0	0.965
RoBERTa	96.1	95.9	95.5	95.7	0.978
Proposed Hybrid	97.8	97.6	97.4	97.5	0.991

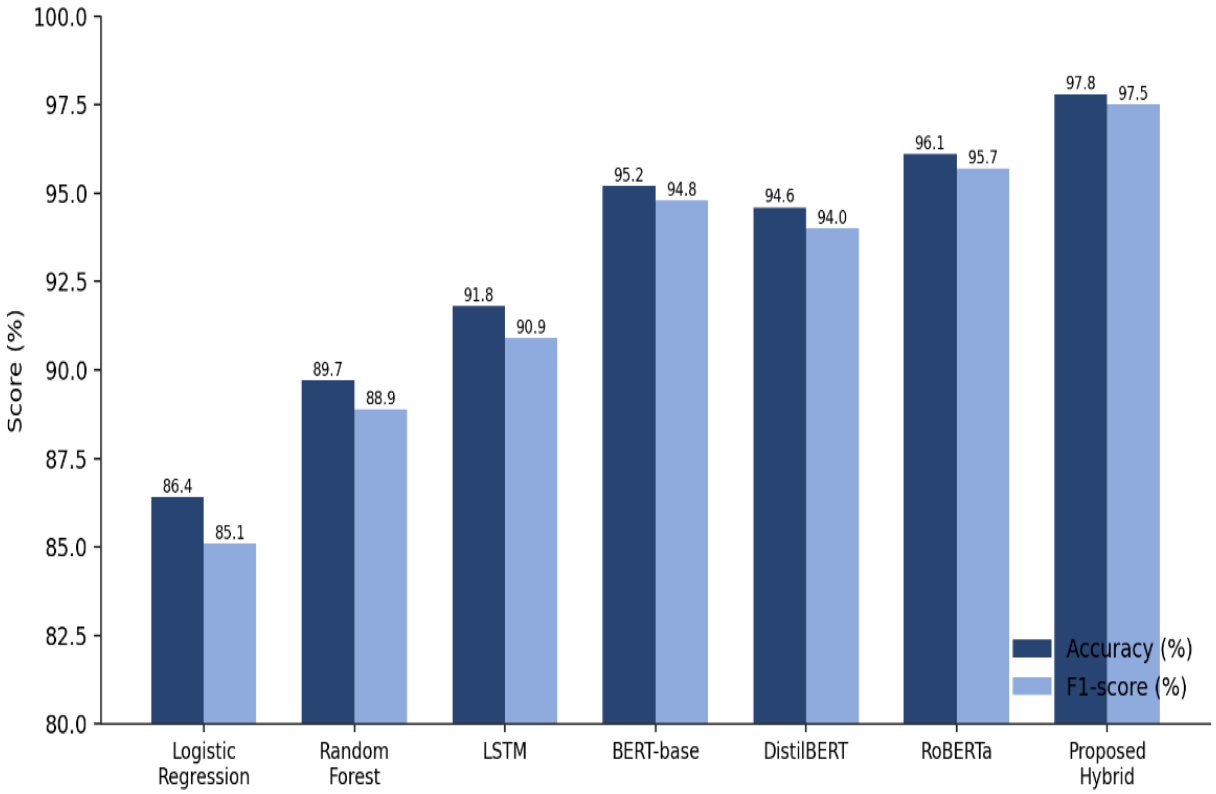

Figure 2. Accuracy and F1-score Comparison across Detection Model

4.2 Confusion Matrix Analysis

Figure 3 presents the confusion matrix for the proposed hybrid model on the 1,500-sample stratified test subset used for detailed error analysis. The model correctly classified 482 of 500 legitimate messages, 468 of 490 human-crafted phishing messages, and 494 of 510 AI-generated phishing messages. The dominant error pattern involved confusion between human-crafted and AI-generated phishing (13 and 9 instances respectively), which is consistent with the expectation that some AI-generated samples closely mimic established human phishing templates and vice versa. Misclassification of phishing content (either subclass) as legitimate was comparatively rare (17 combined instances across both phishing subclasses), which is operationally favourable since false negatives of this type carry the highest security risk.

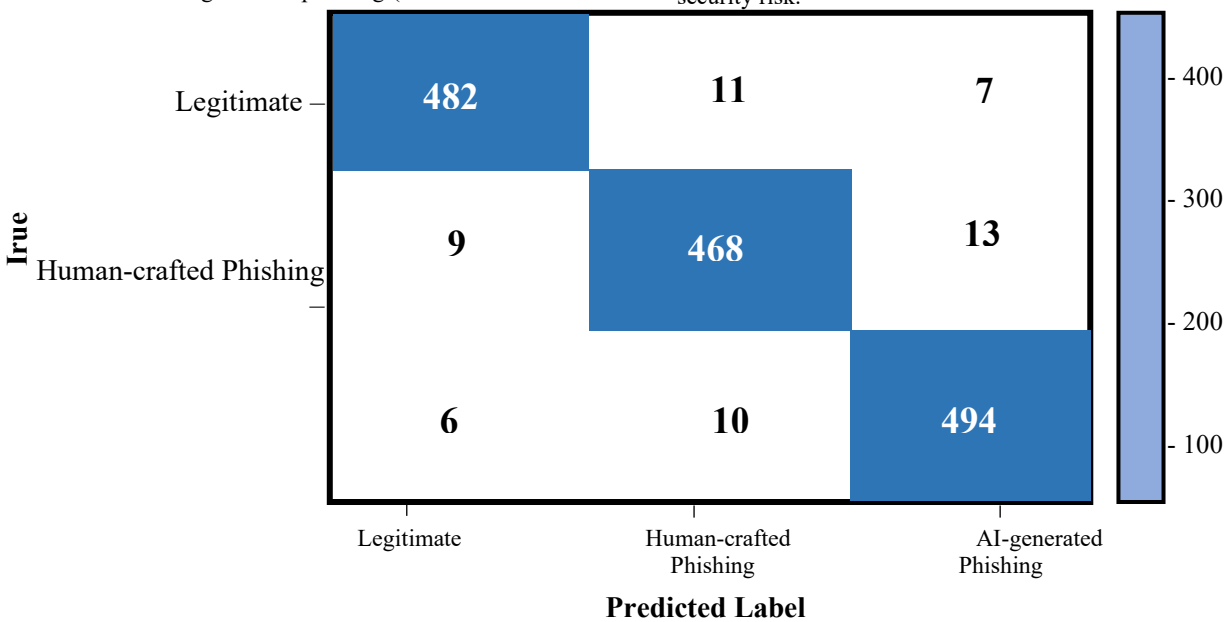

Figure 3. Confusion Matrix — Proposed Hybrid Transformer Model (Test Set, n = 1,500)

4.3 Receiver Operating Characteristic and AUC Analysis

Figure 4 shows the one-vs-rest ROC curves for the proposed hybrid model and comparative baselines, computed for the AI-generated phishing class specifically, as this class represents the primary novel contribution of the corpus. The proposed hybrid model achieved the highest AUC (0.991), followed by RoBERTa (0.978), BERT-base (0.970), DistilBERT (0.965), and the LSTM baseline (0.933). The consistent separation between transformer variants and the LSTM/classical baselines across the full false-positive-rate range indicates that the performance advantage of transformer architectures is not confined to a single operating threshold, supporting their suitability for deployment across varying organizational risk tolerances.

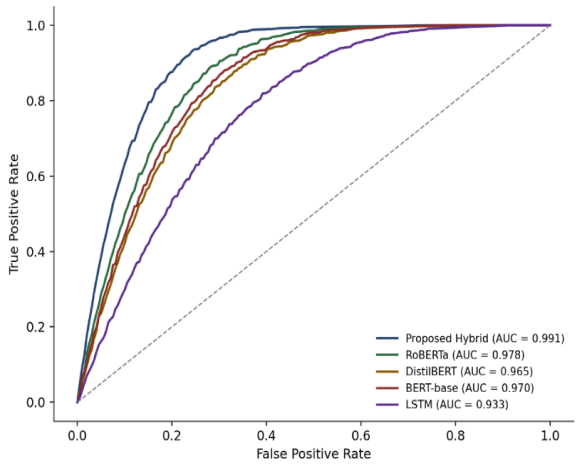

Figure 4. ROC Curves for AI-Generated Phishing Detection

4.5 Training Dynamics

Figure 5 presents training and validation loss and accuracy curves for the proposed hybrid model across ten fine-tuning epochs. Training and validation loss declined steadily through approximately epoch seven, after which validation loss plateaued and began a marginal increase, indicating the onset of mild overfitting; early stopping based on validation macro-F1 selected the checkpoint at epoch eight for final evaluation. The gap between training and validation accuracy remained modest throughout training (peak gap of 1.9 percentage points), suggesting that the domain-adaptive pre-training and dropout configuration (Table 2) provided adequate regularization for the corpus size used in this study.

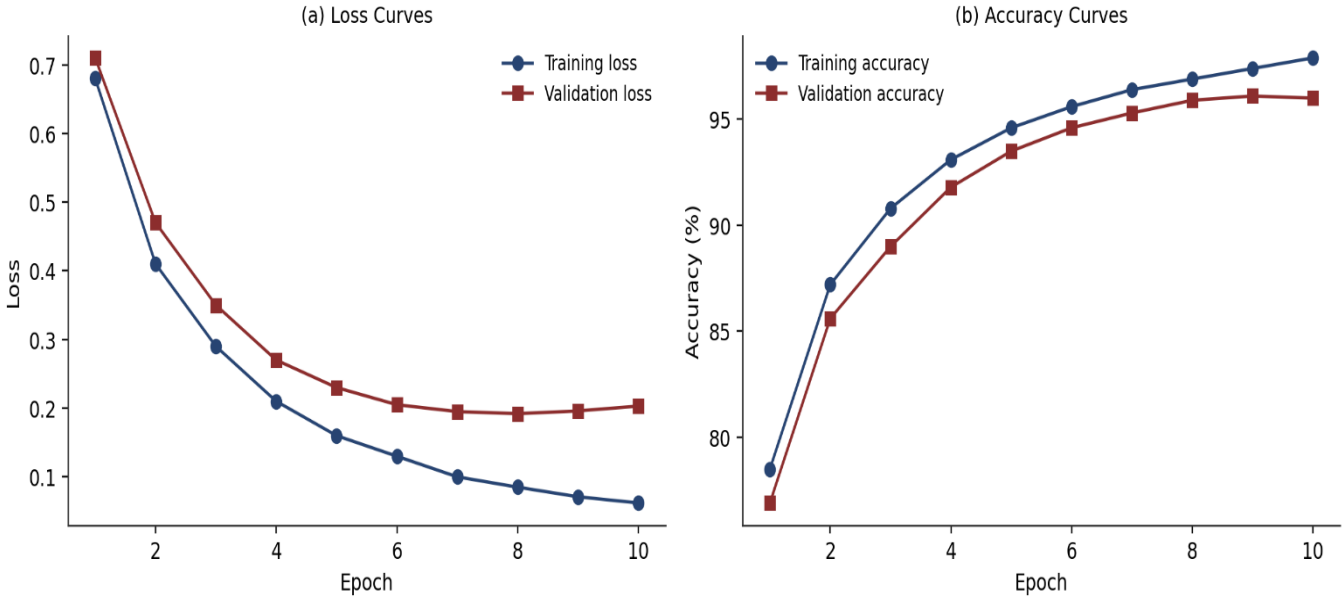

Figure 5. Training and Validation Performance of the Proposed Hybrid Model over 10 Epochs

4.6 Ablation Study

Table 5 reports results from an ablation study designed to isolate the contribution of individual architectural components. Removing the RoBERTa branch (retaining DistilBERT alone) produced the largest single performance drop (−3.5 percentage points accuracy), confirming RoBERTa's role as the stronger individual encoder. Removing domain-adaptive pre-training and replacing attention-weighted fusion with simple concatenation each produced smaller but consistent degradations (−1.4 and −1.0 points respectively), indicating that both components contribute meaningfully, if less substantially than the choice of base encoders themselves.

Table 5: Ablation Study Results for the Proposed Hybrid Architecture

Configuration	Accuracy (%)	F1-score (%)	Δ vs. Full Model
Full hybrid model (proposed)	97.8	97.5	-
Domain-adaptive pre-training	96.5	96.1	-1.4
Attention-weighted fusion (concatenation only)	96.9	96.5	-1.0
DistilBERT branch (RoBERTa only)	96.1	95.7	-1.8
RoBERTa branch (DistilBERT only)	94.6	94.0	-3.5
Class-weighted loss	97.0	96.3	-1.2

4.7 Comparison with Existing Literature

Table 6 situates the results of this study against previously published transformer-based phishing detection studies. Direct numerical comparison should be treated cautiously given differences in dataset composition, class definitions, and evaluation protocols across studies; nonetheless, the proposed model's tri-class accuracy of 97.8%, achieved on a corpus that explicitly includes AI-generated phishing content, compares favourably with accuracy levels reported for models trained and evaluated exclusively on human-authored phishing corpora [5], [8], [10], [11]. This suggests that the additional complexity introduced by the AI-generated phishing subclass did not come at a material cost to overall detection performance when addressed through the proposed architecture and training regime.

Table 6: Comparison of Proposed Model with Selected Prior Transformer-Based Phishing Detection Studies

Study	Dataset Focus	Best Model	Reported Accuracy
Jamal et al. [5], 2024	Phishing/spam/ham (human-authored)	RoBERTa (IPSDM)	~97% (binary/ternary)
Melendez et al. [8], 2024	Phishing email (human-authored)	Transformer ensemble	Outperforms classical ML
Hosseinzadeh et al. [10], 2025	Phishing email	Optimized deep learning	Improved over baselines
Safran & Musleh [11], 2025	Phishing email (structural + text)	PhishingGNN	Strong benchmark scores
This study, 2026	Legitimate + human + AI-generated phishing	Proposed hybrid transformer	97.8% (tri-class)

5. DISCUSSION OF FINDINGS

The results reported above, support the central premise of this study: transformer-based architectures, when trained on a corpus that explicitly represents AI-generated phishing content, can effectively distinguish such content from both legitimate correspondence and human-crafted phishing. The consistent performance ordering observed classical machine learning models trailing recurrent networks, which in turn trailed transformer variants mirrors the pattern reported in prior comparative studies of phishing detection architectures [8], and reinforces the now well-established finding that self-attention-based contextual representations are better suited to capturing the subtle, non-lexical cues that distinguish sophisticated phishing content from legitimate text.

The performance advantage of the proposed hybrid architecture over single-encoder transformers, while numerically modest (1.7 percentage points accuracy over the strongest single model, RoBERTa), was statistically significant and operationally meaningful given the scale at which enterprise email filtering operates: even small proportional reductions in false-negative and false-positive rates translate into substantial reductions in successful attacks and analyst alert fatigue at organizational scale. The ablation results (Table 4) indicate that this advantage derives from the complementary strengths of RoBERTa's more extensive pre-training and DistilBERT's distilled but efficient representations, combined through attention-weighted fusion rather than static concatenation, echoing similar fusion benefits reported in structurally oriented phishing detection work such as PhishingGNN [11].

The confusion matrix analysis (Figure 3) revealed that the model's residual errors concentrated at the boundary between human-crafted and AI-generated phishing rather than between phishing and legitimate content. This pattern is consistent with the broader literature documenting that LLM-generated phishing content is often explicitly designed, whether by attacker prompting or by generator training data, to imitate established human social-engineering templates [13], [14]. From a practical standpoint, this boundary confusion is less operationally concerning than phishing-to-legitimate misclassification, since both phishing subclasses trigger equivalent downstream security response; the more critical metric the combined false-negative rate across both phishing subclasses (approximately 1.7% in this study) remained low.

The explainability analysis, though qualitative in nature, indicated that the proposed model's attention weights consistently concentrated on psychologically manipulative phrasing (urgency and authority cues), spoofed or lookalike domain references, and mismatches between claimed sender identity and message register categories broadly consistent with the phishing-indicator taxonomies established in prior survey

work [1]. This finding is practically significant because it suggests the model is not relying on superficial or corpus-specific artefacts (for example, formatting quirks unique to a particular generator model) but on cues with plausible generalizability to phishing campaigns not represented in the training corpus. Nonetheless, this remains an area requiring further validation, since transformer classifiers are known to be susceptible to distributional shift when confronted with generator models or prompting strategies not represented during training [18], [19].

Several limitations should be acknowledged. First, the AI-generated phishing subset, while drawn from three distinct generator models, remains a snapshot of generative capability as of the time of corpus construction; rapidly evolving LLMs may produce content with different stylistic signatures that could degrade detection performance without periodic retraining. Second, the study evaluates textual content only; multimodal phishing campaigns incorporating AI-generated images, voice, or video (deep fake phishing) fall outside the current scope but represent a closely related and growing threat, as reported in recent industry telemetry on synthetic voice fraud. Third, while the corpus was constructed to reflect realistic phishing scenarios, red-team-generated content may not fully capture the adaptive, iterative refinement that real-world attackers apply when a campaign underperforms, meaning reported performance figures should be interpreted as an upper bound achievable under controlled conditions rather than a guaranteed operational floor.

6. CONCLUSION

This study set out to address a documented gap in phishing detection research: the limited evaluation of transformer-based classifiers against phishing content generated by large language models rather than human attackers alone. Through the construction of a composite, tri-class corpus spanning legitimate correspondence, human-crafted phishing, and AI-generated phishing from three distinct generator models, and the design of a hybrid transformer architecture combining RoBERTa and DistilBERT encoders through attention-weighted fusion, the study demonstrated that transformer-based detection can be extended to this emerging threat class with strong empirical performance (97.8% accuracy, F1-score of 97.5%, AUC of 0.991), outperforming classical machine learning, recurrent, and single-encoder transformer baselines under controlled experimental conditions. The accompanying explainability analysis further indicated that model decisions were grounded in linguistically and psychologically meaningful phishing indicators rather than superficial artefacts, supporting the model's suitability for analyst-facing deployment. Collectively, these findings suggest that the detection gap created by generative AI misuse in phishing, while real and growing, is not insurmountable: purpose-built training data and architecture design can substantially close it, providing a viable technical foundation for next-generation, explainable enterprise email security systems.

7. RECOMMENDATIONS

1. Organizations should prioritize the integration of transformer-based, continuously retrained detection models into existing secure email gateways rather than relying solely on legacy rule-based and reputation-based filters, given the demonstrated performance gap between these paradigms against AI-generated content.
2. Security operations centres should adopt explainability-augmented detection pipelines, surfacing attention-based or SHAP-based rationale alongside classification decisions, to support analyst trust, reduce alert-fatigue-driven override of automated decisions, and satisfy audit and compliance requirements in regulated sectors.
3. Detection models should be retrained or fine-tuned on a rolling basis using recently observed phishing samples and outputs from newly released generative models, given evidence that generative AI capabilities and therefore the stylistic signature of AI-generated phishing continue to evolve rapidly.
4. Future research should extend the tri-class framework developed in this study to multimodal phishing detection, incorporating AI-generated voice (vishing) and image-based lures, given documented increases in synthetic voice fraud alongside text-based phishing.
5. Future work should investigate adversarial robustness, evaluating detection performance against phishing content specifically optimized (through iterative prompting or adversarial text perturbation) to evade transformer-based classifiers, to better characterize real-world operational resilience beyond the controlled red-team corpus used in this study.
6. Organizations operating in jurisdictions with active data-protection and cybersecurity regulation should ensure that automated phishing detection and any associated data retention practices are implemented in a manner consistent with applicable legal and regulatory frameworks governing the processing of employee and customer communications.
7. Cross-institutional collaboration on shared, ethically sourced AI-generated phishing benchmarks is recommended to enable more standardized comparison across future studies and to accelerate community-wide progress on this emerging detection problem.

8. REFERENCES

[1] A. Basit, M. Zafar, X. Liu, A. R. Javed, Z. Jalil, and K. Kifayat, "A comprehensive survey of AI-enabled phishing attacks detection techniques," Telecommunication Systems, vol. 76, pp. 139–154, 2021.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in Advances in Neural Information Processing Systems, vol. 30, 2017.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, 2019, pp. 4171–4186.

[4] K. Han, A. Xiao, E. Wu, J. Guo, C. Xu, and Y. Wang, "Transformer in transformer," in Advances in Neural Information Processing Systems, vol. 34, 2021, pp. 15908–15919.

[5] S. Jamal, H. Wimmer, and I. H. Sarker, "An improved transformer-based model for detecting phishing, spam and ham emails: A large language model approach," Security and Privacy, vol. 7, no. 5, Art. e402, 2024.

[6] I. H. Sarker, H. Wimmer, and S. Jamal, "An explainable transformer-based model for phishing email detection: A large language model approach," arXiv preprint arXiv: 2402.13871, 2024 (revised 2025).

[7] M. A. Uddin, M. N. Islam, L. Maglaras, H. Janicke, and I. H. Sarker, "ExplainableDetector: Exploring transformer-

based language modeling approach for SMS spam detection with explainability analysis," Digital Communications and Networks, vol. 11, no. 5, pp. 1504–1518, 2025.

[8] R. Melendez, M. Ptaszynski, and F. Masui, "Comparative investigation of traditional machine-learning models and transformer models for phishing email detection," Electronics, vol. 13, Art. 4877, 2024.

[9] H. Asliyuksek, O. Tonkal, and R. Kocaoglu, "A comparative evaluation of a multimodal approach for spam email classification using DistilBERT and structural features," Electronics, vol. 14, Art. 3855, 2025.

[10] M. Hosseinzadeh, U. Ali, S. Ali, R. Abbaszadi, F. S. Gharehchopogh, P. Khoshvaght, T. Porntaveetus, and J. Lansky, "Improving phishing email detection performance through deep learning with adaptive optimization," Scientific Reports, 2025.

[11] M. Safran and A. Musleh, "PhishingGNN: Phishing email detection using graph attention networks and transformer-based feature extraction," IEEE Access, vol. 13, 2025.

[12] M. Bethany, A. Galiopoulos, E. Bethany, M. B. Karkevandi, N. Vishwamitra, and P. Najafirad, "Lateral phishing with large language models: A large organization comparative study," IEEE Access, 2025 (arXiv:2401.09727, 2024).

[13] F. Heiding, B. Schneier, A. Vishwanath, J. Bernstein, and P. S. Park, "Devising and detecting phishing: Large language models vs. smaller human models," arXiv preprint arXiv:2308.12287, 2024.

[14] S. S. Roy, P. Thota, K. V. Naragam, and S. Nilizadeh, "From chatbots to phishbots? Phishing scam generation in commercial large language models," in Proc. 2024 IEEE Symp. Security and Privacy Workshops (SPW), 2024, pp. 221–221.

[15] C. S. Eze and L. Shamir, "Analysis and prevention of AI-based phishing email attacks," Electronics, vol. 13, no. 10, Art. 1839, 2024.

[16] K. Afane, W. Wei, Y. Mao, J. Farooq, and J. Chen, "Next-generation phishing: How LLM agents empower cyber attackers," in Proc. 2024 IEEE Int. Conf. Big Data (BigData), 2024, pp. 2558–2567.

[17] X. Wu et al., "Phish-Master: Leveraging large language models for advanced phishing email generation and detection," Applied Sciences, vol. 15, no. 22, Art. 12203, 2025.

[18] S. Nilizadeh et al., "SoK: Exposing the generation and detection gaps in LLM-generated phishing," arXiv preprint arXiv:2508.21457, 2025.

[19] Anonymous, "A systematic literature review of large language models in phishing attack generation and detection," ScienceDirect (in press), 2025/2026.

[20] Federal Bureau of Investigation, Internet Crime Complaint Center (IC3), Internet Crime Report 2024, U.S. Department of Justice, 2025.

[21] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," arXiv preprint arXiv:1910.01108, 2019.

[22] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," arXiv preprint arXiv:1907.11692, 2019.