2-in-1 Phishing Detection via Large LM Distillation and Small LM Perturbation (Student Abstract)

Calvin Greenewald, Bradley Ashmore, Chien-Sing Poon, Lingwei Chen

Wright State University, Dayton, OH 45435, USA
{greenewald.6,ashmore.3,chien.poon,lingwei.chen}@wright.edu

Abstract

Phishing emails are an escalating threat, underscoring the need for precise detection methods. While large language models (LLMs) have gained attention for their potential in this area, their reliance on extensive data for fine-tuning poses practical challenges. This paper introduces DualLM for phishing detection with minimal data, which distills the reasoning ability from a large LM to enhance a small target LM and integrates trainable perturbations to improve the small LM's inference capabilities. Experiments demonstrate that DualLM can benefit from dual LMs, which reduces training parameters and data required, while maintaining high performance in phishing email detection with limited data.

Introduction

Emails are prime targets for phishing, and deep learning is commonly used for phishing detection, but it requires large labeled datasets. Due to privacy concerns and high annotation costs, acquiring sufficient data is challenging. While transfer learning with LLMs offers adaptability, fine-tuning still demands substantial labeled data. Recent work on LLM distillation into small LMs reduces data needs (Hsieh, Li, and et al. 2023), but full-layer optimization still induces large parameters, limiting their use in data-limited settings.

In this paper, we propose DualLM, a simple and efficient mechanism for phishing detection with limited data. DualLM uses a large LM to generate rationales from labeled emails (Kojima et al. 2022), guiding the reasoning of a small target LM. To transfer the small LM for phishing detection, we freeze its parameters and train only the head, which offers stability with small data but lacks inference effectiveness. To address this, trainable perturbations are appended to inputs and propagated throughout the small LM, where non-linear self-attention mechanism enables these perturbations to act as parameters that introduce variability into the frozen layers. DualLM reduces training parameters and data needs while maintaining promising detection performance.

Design of DualLM

The overview of DualLM is depicted in Figure 1: a large LM generates rationales to guide a small target LM's reasoning, which further reduces data needs via input perturbations.

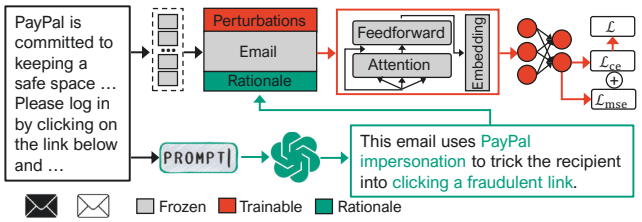

Figure 1: The overview of DualLM.

Task	Prompt Template
Binary or Multi-class	The email is [content] with categories $[c_1, ..., c_k]$. Given its true class is $[c_i]$, predict the label and provide a detailed explanation of why.

Email	Generated Rationale
CONGRATS! You Can Get $50 Walgreens Rewards. CLICK HERE To See it!	The email employs a sense of urgency and an unsolicited reward indicating phishing.

Table 1: Prompt template and rationale example.

Reasoning LLM for Rationale Extraction

We employ the chain-of-thought prompting (Wei et al. 2022) for this purpose. A prompt template is first curated to guide the LLM in articulating how a labeled email $X_i \in \mathcal{X}_l$ can be classified as $y_i \in \mathcal{Y}$. The template specifies the email X_i, its ground truth y_i, and the chain-of-thought prompt P. The design of our prompt template is illustrated in Table 1. Using the curated prompt template, we populate each input X_i and its ground truth y_i into the template, and use the formulation to prompt the LLM and extract the rationale R_i for each labeled email. This process can be written as:

$$R_i = \text{LLM}(X_i, y_i, P),\ X_i \in \mathcal{X}_l,\ y_i \in \mathcal{Y} \qquad (1)$$

Table 1 provide an email with its rationale by prompting GPT-3.5-Turbo, clearly explaining why it is phishing.

Perturbing Small LM for Phishing Detection

To transfer the small LM to the target task, a simple method to reduce parameters is to freeze the model and train only the head, but this limits task-specific learning. To overcome this, we introduce trainable perturbations that interact non-linearly with input tokens via self-attention to adapt the

Models	Phishing-Type	Phishing-Spam	Phishing-Fraud
BERT-F	14.88 ± 0.71	65.15 ± 0.39	38.92 ± 0.30
RoBERTa-F	27.74 ± 0.82	60.76 ± 0.84	38.92 ± 0.40
DistilBERT-F	16.57 ± 0.97	87.77 ± 0.10	90.56 ± 0.51
BERT-H	61.37 ± 0.81	85.77 ± 0.34	88.31 ± 0.25
RoBERTa-H	60.88 ± 0.83	84.87 ± 0.84	92.50 ± 0.38
DistilBERT-H	68.08 ± 0.58	90.78 ± 0.10	91.22 ± 0.20
Distilling	61.88 ± 0.59	88.95 ± 0.17	92.54 ± 0.20
WARP	63.46 ± 0.81	91.09 ± 0.10	92.92 ± 0.14
DualLM-BERT	**74.87 ± 0.93**	93.43 ± 0.20	96.52 ± 0.08
DualLM-RoBERTa	68.40 ± 0.82	**94.04 ± 0.18**	97.20 ± 0.08
DualLM-DistilBERT	74.85 ± 0.79	90.89 ± 0.20	**97.31 ± 0.09**

Table 2: Comparison with Baselines (F1 Score %).

frozen layers and improve performance. Formally, these perturbations are defined as $\widehat{\mathbf{\Theta}} = \{\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_t\}$. Given an email $\mathbf{X}_i = \{\mathbf{x}_1, ..., \mathbf{x}_m\}$ and its rationale $\mathbf{R}_i = \{\mathbf{r}_1, ..., \mathbf{r}_l\}$, the new input sequence for the small LM can be formulated as:

$$\mathbf{X}_i^s = \{[cls], \mathbf{r}_1, \cdots, \mathbf{r}_l, \mathbf{x}_1, \cdots, \mathbf{x}_m, \boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_t, [sep]\} \tag{2}$$

where rationale guides phishing-specific focus, and perturbations enhance variability. This reduces trainable parameters to $t \times 768$ (with t expected small) in addition to the head, achieving up to $10,000$ times fewer parameters. The non-linear interactions in $\mathbf{X}_i^s$ are essential that can be regulated by the LM's multi-head self-attention mechanism.

Optimization

The representation $\mathbf{H}_i$ at $[cls]$ is fed to the head, producing $\mathbf{Z}_i$=softmax$(f_{\widetilde{\mathbf{\Theta}}}(\mathbf{H}_i))$ for $\mathbf{X}_i$. The total parameter set for DualLM is $\mathbf{\Theta} = \{\widehat{\mathbf{\Theta}}; \widetilde{\mathbf{\Theta}}\}$. Optimization involves two objectives: a cross-entropy loss

$$\mathcal{L}_{ce} = -\frac{1}{|\mathcal{X}_l|} \sum_{X_i \in \mathcal{X}_l} y_i \log \mathbf{Z}_{i_{y_i}} \tag{3}$$

to measure prediction correctness, and a mean squared loss

$$\mathcal{L}_{mse} = \frac{1}{|\mathcal{X}_l|} \sum_{X_i \in \mathcal{X}_l} (\mathbf{H}_i - \text{max-pooling}(\mathbf{R}_i))^2 \tag{4}$$

to align $\mathbf{H}_i$ with the rationale $\mathbf{R}_i$. The final loss is $\mathcal{L} = \mathcal{L}_{ce} + \lambda \mathcal{L}_{mse}$ where λ is a balance parameter. During inference, as label is unavailable, only the trained perturbations are used, without rationale generation.

Evaluation

We evaluate DualLM on three email corpora: Phishing-Type (Hall 2024) (160 emails with 4 classes), Phishing-Spam (Jackksoncsie 2024) (5,685 emails with 2 classes), and Phishing-Fraud (Verma 2024) (12,000 emails with 2 classes). We randomly select n-shot samples for training, 20% for testing, and 20% for validation. Some closely related baselines are selected for comparison: 3 relatively small LMs: BERT, RoBERTa, and DistilBERT (two variants for each: X-F for fine-tuning and X-H for head-training), Distilling Step-by-Step (Hsieh, Li, and et al. 2023), WARP (Hambardzumyan and et al. 2021). We use GPT-3.5 Turbo as LLM, and set $\lambda = 0.5$.

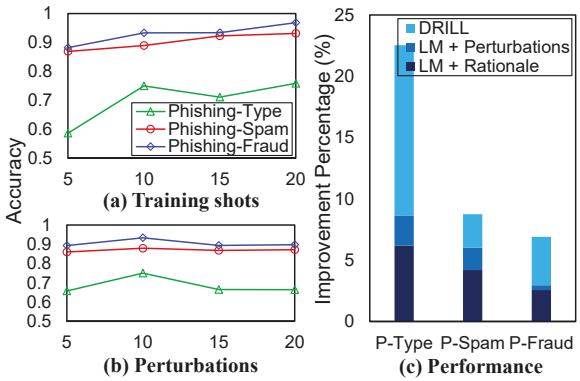

Figure 2: (a) Training shots, (b) Perturbation numbers, and (c) performance improvement by different components.

Email	**Rationale**
Hi, we're offering a student empowerment program for part-time jobs with attractive wages. **Reply with your details** if interested.	The email shows **phishing** signs by **requesting personal details for malicious use**.

Model	True Label	Predicted Label
BERT	Phishing	Legitimate
Rationale Distillation	Phishing	Legitimate
DualLM	Phishing	Phishing

Table 3: Case study.

Results

We compare DualLM with baselines and report the results ($n = 10$ and $t = 10$) in Table 2. DualLM improves the performance upon BERT, RoBERTA, and DistilBERT and also outperforms Distilling and WARP that leverages rationales and input token perturbations, respectively, demonstrating the benefit of dual-LM scheme. We also evaluate the impact of training shots, perturbations, and different components. As shown in Figure 2(a) and (b), performance improves with increasing n, while a larger t does not always lead to better results. Figure 2(c) indicates that adding rationale and perturbations significantly improves performance over the BERT baseline, with perturbations contributing more.

Case Study

Table 3 provides an email example and its rationale, as well as the prediction results from BERT, rationale distillation, and DualLM. Our model successfully detects this email as phishing by leveraging both email semantics via perturbations and rationales for the enhancement, reaffirming the importance of dual-LM operations.

Conclusion

In this paper, we propose DualLM for data-limited phishing detection, which distills the reasoning ability from an LLM into a target small LM, and integrates trainable perturbations to inputs to enhances its inference ability. Extensive experimental results demonstrate that DualLM reduces data required and maintains promising detection performance.

Acknowledgments

This work is partially supported by the NSF under grant CNS-2245968.

References

Hall, C. 2024. Phishing Email Data by Type. https://www.kaggle.com/datasets/charlottehall/phishing-email-data-by-type. Accessed: 2024-12-05.

Hambardzumyan, K.; and et al. 2021. Warp: Word-level adversarial reprogramming. *arXiv preprint arXiv:2101.00121*.

Hsieh, C.-Y.; Li, C.-L.; and et al. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *ACL*.

Jackksoncsie. 2024. Spam Email Dataset. https://www.kaggle.com/datasets/jackksoncsie/spam-email-dataset. Accessed: 2024-12-05.

Kojima, T.; Gu, S. S.; Reid, M.; Matsuo, Y.; and Iwasawa, Y. 2022. Large language models are zero-shot reasoners. *NeurIPS*, 35: 22199–22213.

Verma, A. 2024. Fraud Email Dataset. https://www.kaggle.com/datasets/llabhishekll/fraud-email-dataset. Accessed: 2024-12-05.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS*, 35: 24824–24837.