DRILL: Dual-Reasoning Large Language Models for Phishing Email Detection with Limited Data

Calvin Greenewald, Bradley Ashmore, Chien-Sing Poon, and Lingwei Chen(✉)

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

Abstract. As phishing emails pose a growing threat to individuals and organizations alike, there is an urgent need to develop more accurate detection methods. Large Language Models (LLMs) have recently garnered major attention in this line of research; however, they often require large-scale data for fine-tuning, which is impractical in real-world application scenarios. This paper proposes DRILL, a new simple and efficient mechanism, for **D**ual-**R**eason**I**ng **LL**Ms to detect phishing emails with extremely small data. DRILL distills the reasoning ability from an LLM into a target small LM model, while integrating trainable perturbations to manipulate the inputs, which in turn adaptively enhances the inference ability of the target LM. Extensive experiments are conducted on multiple real-world email datasets, and the evaluation results demonstrate that DRILL can benefit from dual LMs, which significantly reduces training parameters and data required, while maintaining state-of-the-art performance in phishing email detection with limited data.

Keywords: Phishing Email Detection · Large Language Models · Reasoning · Data-Limited Learning.

1 Introduction

Emails serve as a ubiquitous and regular form of communication for both individuals and organizations, making them a primary target for fraudulent activities [33]. Cybercriminals have been increasingly exploiting email vulnerabilities through a surge of phishing attacks in attempts to gain access to personal information, financial assets, or install malware onto legitimate systems [34, 9, 3, 16]. Due to recent developments in generative AI for text generation [5], not only have phishing emails shot up many-fold [16], but their formulations are also more realistic and grammatically correct [13], making them more challenging to detect [35]. As such, deep learning (DL), which competes with human intelligence, has emerged as a primary method for detecting new phishing emails far earlier than traditional rule-based or signature-based approaches [2, 1, 29, 39, 10, 38]. For example, Zavrak et al. [39] combined CNNs and GRUs with attention to highlight email semantics, and enhance detection effectiveness. Alhogail et al. [1] constructed a word co-occurrence graph to represent the email corpus and reduced email detection to a node classification problem using GCNs.

Despite their effectiveness and scalability, these DL-based detection models rely heavily on large-scale labeled email datasets, which may lead to poor generalization to unseen data and evolving phishing techniques when only limited data is available [24]. In real-world application scenarios, acquiring a sufficient quantity of labeled email samples is challenging due to constraints on data acquisition for privacy concerns and the high cost of annotation. To address data-limited learning challenge, transfer learning [42] is often regarded as a promising approach that repurposes a pretrained model to operate on a target task. Large language models (LLMs), in particular, have recently garnered major attention in transfer learning due to their versatility and adaptability [8, 32, 27, 4]. These models provide powerful learning capabilities to extract intricate semantic patterns from input sequences, enabling them to excel at various tasks, such as phishing email detection [22, 16]. Unfortunately, LLMs are characterized by extensive parameters (e.g., GPT-3 [4] is made up of 175 billion parameters). The prevalent strategy to fine-tune these parameters still requires a substantial amount of labeled samples to yield good results. To reduce the need for training samples, recent works have explored various transfer learning solutions on smaller LMs by generating the reasoning information from LLMs to either augment input prompts for target LMs [37, 23], serve as additional fine-tuning data to enlarge training corpus [40, 19], or perform reasoning distillation to enforce target LMs to emulate the LLMs [12, 18, 17, 26, 28]. These approaches improve the performance of target LMs with less training data, but suffer from a significant drawback: regardless of how inputs and optimization problems are refined or what "small" LMs are trained for the target tasks, model fine-tuning throughout all layers remains unchanged; this still induces large parameters and necessitating relatively large task-specific samples, thus restricting their utility in extremely data-limited applications. This naturally poses a research question: "*Can the reasoning ability of LLMs be transferred to small LMs with significantly fewer trainable parameters, making them better suited for data-limited tasks?*"

In this paper, we propose DRILL, a new simple and computationally efficient mechanism, for **D**ual-**R**eason**I**ng **LL**Ms to detect phishing emails with extremely small data. DRILL proceeds by distilling the reasoning ability from an LLM into our target model, a much smaller LM, while simultaneously integrating the generated rationales with trainable perturbations to manipulate the inputs, which in turn adaptively enhances the reasoning ability of the target small LM, enabling it to capture subtle semantics indicative of phishing intents and achieve high performance in phishing email detection. More specifically, given their demonstrated zero-shot and emergent learning capabilities for open-ended tasks [7, 23, 41, 36], we first leverage an LLM to perform on the limited labeled emails and generate high-quality rationales that justify its predictions to supervise the reasoning of a target small LM. To transfer our target small LM for phishing email detection, we freeze its parameters and exclusively train a task-specific head rather than model fine-tuning. This suboptimal paradigm may offer better stability for small data, but lack satisfactory inference effectiveness due to its shallow learning nature [15]. As such, trainable perturbations are

further appended to the inputs and propagated throughout the inner structure of LM, where the intrinsic nonlinearity from self-attention mechanism allows these perturbations to act as parameters that introduce variability into the frozen layers, modulating how the inputs are processed by the LM. Such a formulation in DRILL can benefit from dual LMs, which significantly reduces the amount of training parameters and data required, while maintaining promising detection performance. In summary, this paper has the following major contributions:

- We cast phishing email detection as a learning problem to transfer the reasoning and inference capabilities on dual language models.
- We design a simple yet effective mechanism DRILL that leverages rationale distillation and trainable perturbations within inputs to significantly reduce training parameters and data required.
- Extensive experiments are conducted on real-world email datasets, demonstrating our model can achieve state-of-the-art phishing email detection performance with limited data.

2 Preliminaries

Notations We define an email corpus as $\mathcal{D} = (\mathcal{X}, \mathbf{y})$, where $X_i \in \mathcal{X}$ represents an individual email, and each labeled sample X_i carries a ground truth denoted by $y_i \in \{0 : \text{legitimate}, 1 : \text{phishing}\}$. We tokenize each email to a set of word tokens, convert each token into a d-dimensional vector, and accordingly map the email corpus $\mathcal{X}$ into the embedding space $\phi : \mathcal{X} \to \mathbf{X} \in \mathbb{R}^{n \times m \times d}$, where n is the number of emails, m is the number of tokens within each email, and d is the dimension of the embedding space.

Data-Limited Phishing Email Detection Due to high cost of annotation and limited access to comprehensive email information, large-scale labeled email data is unlikely to be available in practice. In this respect, we practically consider scenarios where only a limited number of the emails have labels. To emulate real-world constraints, the email corpus $\mathcal{X}$ is divided into two different sets: (1) labeled email set $\mathcal{X}_l$ and (2) unlabeled email set $\mathcal{X}_u$, where $|\mathcal{X}_l| \ll |\mathcal{X}|$. A data-limited phishing email detection problem is then cast as a text classification problem, which uses few labeled examples to learn a model $f_{\boldsymbol{\Theta}} : \mathbf{X} \to \mathbf{y}$ that can effectively predict the labels for unlabeled emails from $\mathcal{X}_u$.

3 Proposed Model

In this section, we present our proposed model DRILL, which leverages dual-reasoning LLMs to enhance the performance of phishing email detection with limited data. Its overview is illustrated in Figure 1. Specifically, two LMs are involved in DRILL: an LLM is used to reason about its predictions and serve as the "teacher" model to generate high-quality rationales and supervise the training

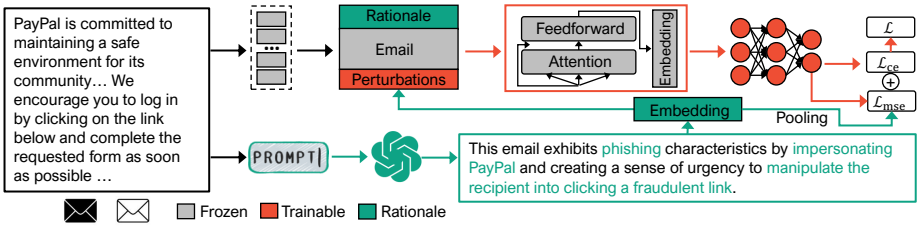

Fig. 1: The overview of our proposed model DRILL by dual-reasoning LLMs for data-limited phishing email detection.

of the target model in a data-efficient way; a small LM is used to optimize its feature learning and act as the target "student" model to perform phishing email detection that further reduces the need for training email samples.

3.1 Reasoning LLM for Rationale Extraction

Since the zero-shot reasoning is known as an emergent ability of LLMs [37, 36], we can extract this reasoning information, referred to as rationale, to understand why a particular input email is characterized as a phishing email or belongs to a specific type of phishing attack. In other words, the rationale here provides a detailed explanation that encodes intricate task-specific knowledge, which, however, is often challenging to reason from small LMs on their own with less sophistication and complexity, particularly when available data is scarce. Intuitively, distilling rationales extracted from LLMs into small LMs may bridge this reasoning gap by reinforcing the small models to approximate the behaviors of LLMs, and enable small LMs to perform better in downstream tasks with significantly less training effort and data. As such, we propose to access LLMs to elicit proper rationales that correspond to input emails and their truth labels.

Prompt Curation We employ the chain-of-thought prompting [37, 23] to guide LLMs in generating rationales. Specifically, a prompt template is first meticulously curated to instruct the LLM on articulating how a given raw labeled email $X_i \in \mathcal{X}_l$ can be identified as phishing, legitimate, or a specific phishing type $y_i \in \mathcal{Y}$. The template specifies the email X_i, its ground truth y_i, and the chain-of-thought prompt P to stimulate the LLM's reasoning process. The design of our prompt template is illustrated in Table 1, which provides consistency in rationale extraction and ensures that the generated rationales are comprehensive and insightful for email classification decisions.

Rationale Extraction Using the curated prompt template, we populate each input X_i and its ground truth y_i into the template, and use the resulting formulation P_i to prompt the LLMs through black-box APIs and extract the rationale

Table 1: Prompt template and rationale example

Task	**Prompt Template**
Binary or Multi-class	The email is [*raw content*] and belongs to the categories $[c_1, c_2, ..., c_k]$. Given that its true class is [*ground truth*], please predict this email's classification label and provide a detailed explanation of why.

Email	**Generated Rationale**
CONGRATS! You Can Get $50 Walgreens Rewards. We have been trying to reach you. Please respond. CLICK HERE To See it!	The email employs a sense of urgency and an unsolicited reward indicating phishing.
Hello, I am Wes, and we are partners for the upcoming assignments. What would be the best way to reach you so we can pick out our assignment? -Wes	This email is written in a professional and concise manner with a straightforward request for collaboration, indicating a legitimate purpose.

R_i for each labeled email. This process can be written as:

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

Table 1 provides two examples of emails along with their corresponding rationales generated by prompting GPT-3.5-Turbo. It is evident that these rationales elucidate the reasoning behind the classification, offering a clear explanation why the given email is a phishing or legitimate email. For instance, the rationale generated for the phishing email correctly identifies the use of urgency and unsolicited rewards—common tactics in phishing schemes—and highlights how the emphasis on immediate action and unverified rewards serve as strong indicators of phishing intent. Similarly, the rationale generated for the second email supports its classification as legitimate.

3.2 Reasoning Small LM for Target Task

Given the rationales from the LLM represented as $\mathbf{R}$ by tokenizing R into word tokens and mapping them to an r-dimensional embedding space, the small LM can easily leverage them to enrich its input semantics [37, 23], increase sample size [40, 19], or distill reasoning ability [21, 18], which accordingly facilitates the transfer of its reasoning and inference capabilities for phishing email detection. Despite all apparent benefits, fine-tuning the entire small LM still faces limitations due to the large number of parameters that need to be updated. For example, the BERT-base model [8] totals 110 million parameters, making it computationally expensive and data-intensive to fine-tune fully. A straightforward way to reduce trainable parameters is to freeze the model and only train the task-specific head [30], which, however, often fails to effectively adapt the pre-trained model to a new application due to the lack of comprehensive task-specific learning, limiting the model's ability to capture new nuances and complexities. To address this issue, inspired by task-specific prompt-tuning [25] and adversarial reprogramming [11, 6], we introduce trainable perturbations that act as

prompts to interact with the input tokens in a non-linear way through the self-attention structure in the small LM. By propagating through the LM's layers, these perturbations dynamically adjust the way inputs are processed, allowing the behaviors of the frozen layers to be flexibly adapted without updating their parameters directly, thereby enhancing the performance of the downstream task.

Motivation Consider the small LM $f_s(\cdot)$ receiving an input $\mathbf{X}'_i = f_s(\mathbf{X}_i, \boldsymbol{\theta}) = \mathbf{X}_i + \boldsymbol{\theta}$ that integrates a task sample (i.e., a token vector sequence derived from an email) $\mathbf{X}_i$ ($\mathbf{X}_i \in \mathbf{X}$) with the perturbation $\boldsymbol{\theta}$. The nonlinear nature of an attention operation enables $\boldsymbol{\theta}$ to be effectively multiplied with $\mathbf{X}_i$, producing an output that can be approximated as:

$$\text{Attention}(\mathbf{X}_i + \boldsymbol{\theta}) = \text{Comp}_{\mathbf{X}_i} + \text{Comp}_{\boldsymbol{\theta}} + \text{Comp}_{\boldsymbol{\theta}^T \mathbf{X}_i} \tag{2}$$

Here, the perturbation $\boldsymbol{\theta}$ acts as a new parameter that adapts the small LM through the component $\text{Comp}_{\boldsymbol{\theta}^T \mathbf{x}_t}$ without altering the original model parameters. As a result, only the perturbation $\boldsymbol{\theta}$ is updated throughout this process, while all original parameters of the pretrained model remains unchanged. This offers two primary benefits: (1) reducing model complexity with notably smaller number of trainable parameters and lowering the demand for labeled data and computational cost; and (2) utilizing the pretrained LM's powerful learning capabilities to capture intricate semantic patterns from input tokens, potentially achieving high performance in phishing email detection.

Input Reformation Formally, to enable the small LM to perform reasoning, we define a set of universal perturbations to interact with the input tokens as:

$$\widehat{\boldsymbol{\Theta}} = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3, ..., \boldsymbol{\theta}_t\} \tag{3}$$

where t is a hyperparameter determining their quantity. Given an token sequence representing an email $\mathbf{X}_i = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_m\}$ and another token sequence representing its rationale $\mathbf{R}_i = \{\mathbf{r}_1, \mathbf{r}_2, ..., \mathbf{r}_l\}$, the new input sequence $\mathbf{X}_i^s$ for the target small LM can be constructed as follows:

$$\mathbf{X}_i^s = \{[cls], \mathbf{r}_1, \cdots, \mathbf{r}_l, \mathbf{x}_1, \cdots, \mathbf{x}_m, \boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_t, [sep]\} \tag{4}$$

where the rationale encourages the model to focus on phishing-specific semantics in the email, while the perturbations offer a high degree of variability for the LM's adaptation. This design reduces the number of trainable parameters to $t \times 768$ (with t expected to be a small value and 768 as the typical dimension of individual token embedding) in addition to the task-specific classification head, potentially contributing up to $10,000$ times fewer parameters. The impact of t on the model performance will be evaluated in Section 4.2.

Non-linear Interaction To reason the small LM for phishing email detection through trainable perturbations, the non-linear interactions among all tokens

in $\mathbf{X}_i^s$ are essential, which can be implemented by the multi-head self-attention mechanism built into the LM. Specifically, the attention layer learns to dynamically assign attention weights that capture the relevance between the perturbations and each token. The key operations in the multi-head attention layer involve calculating self-attention scores for each head, which can be expressed through the following equations:

$$\text{Attention}\left(\mathbf{Q}, \mathbf{K}, \mathbf{V}\right) = \text{softmax}\left(\mathbf{Q}\mathbf{K}^{\mathbf{T}}/\sqrt{d_k}\right)\mathbf{V} \tag{5}$$

$$\mathbf{Q} = \mathbf{W}^Q\mathbf{X}_i^s,\ \mathbf{K} = \mathbf{W}^K\mathbf{X}_i^s,\ \mathbf{V} = \mathbf{W}^V\mathbf{X}_i^s \tag{6}$$

where $\mathbf{Q}$, $\mathbf{K}$ and $\mathbf{V}$ are matrices derived by applying weight matrices that have been pretrained in the LM to the input $\mathbf{X}_i^s$, and the term $\frac{1}{\sqrt{d_k}}$ acts as a scaling factor to ensure that the dot products do not become too large that would result in overly sharp softmax distributions, destabilizing training. It is evident that the perturbations can be converted into parameters through $\mathbf{Q}\mathbf{K}^{\mathbf{T}}$ to communicate with the LM, leading to its improved reasoning and inference capabilities.

3.3 Optimization and Inference

Following multiple attention layers, the feature representation $\mathbf{H}_i$ at the $[cls]$ position is fed to the head that derives the prediction probability vector $\mathbf{Z}_i = \text{softmax}(f_{\boldsymbol{\Theta}}(\mathbf{H}_i))$ for the input email $\mathbf{X}_i$. This linear transformation further introduces trainable weights $\widetilde{\boldsymbol{\Theta}}$, such that the parameter sets for DRILL to update can be finalized as:

$$\boldsymbol{\Theta} = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3, ..., \boldsymbol{\theta}_t; \widetilde{\boldsymbol{\Theta}}\} \tag{7}$$

The optimization of DRILL is formulated as two objectives: one is for detection, and another one is for rationale distillation.

Loss for Detection DRILL utilizes cross-entropy loss to optimize the detection task, which measures the difference between the predicted probabilities and the ground truth labels as follows:

$$\mathcal{L}_{ce} = -\frac{1}{|\mathcal{X}_l|}\sum_{X_i \in \mathcal{X}_l} y_i \log \mathbf{Z}_{i_{y_i}} \tag{8}$$

where $\mathbf{Z}_{i_{y_i}}$ is the probability that predicts the email X_i as class y_i.

Loss for Rationale Distillation To avoid introducing extra parameters for rationale generation, DRILL simply aligns the representation $\mathbf{H}_i$ from the small LM with the rationale $\mathbf{R}_i$ extracted from the LLM by assessing a mean squared error between them. Specifically, we first utilize max-pooling to aggregate $\mathbf{R}_i$ to a single representation vector, and then calculate the loss as follows:

$$\mathcal{L}_{mse} = \frac{1}{|\mathcal{X}_l|}\sum_{X_i \in \mathcal{X}_l} \|\mathbf{H}_i - \text{max-pooling}(\mathbf{R}_i)\|^2 \tag{9}$$

Table 2: Statistics of the datasets (note: we evaluate DRILL using n-shot settings, where for training samples, $n \in \{5, 10, 15, 20\}$)

Dataset	**#Email**	**#Class**	**#Training**	**#Validation**	**#Test**
Phishing-Type	160	4	$n \times 4$	32	32
Phishing-Spam	5,685	2	$n \times 2$	1,137	1,137
Phishing-Fraud	12,000	2	$n \times 2$	2,400	2,400

The final loss function for optimizing DRILL is

$$\mathcal{L} = \mathcal{L}_{ce} + \lambda \mathcal{L}_{mse} \tag{10}$$

where λ is a balance parameter to trade off $\mathcal{L}_{ce}$ and $\mathcal{L}_{mse}$. This combination allows DRILL to simultaneously transfer dual-reasoning abilities from both models, leading to more effective inference with much less training data.

Inference For inference on unlabeled (or test) emails $\mathcal{X}_u$, we only integrate the emails with the trained perturbations and feed them into the target small LM for classification. Since no ground truth label is available, we do not generate rationales in this process.

4 Experimental Results and Analysis

4.1 Experimental Setup

Datasets We evaluate DRILL on three real-world email corpora: Phishing-Type [14], Phishing-Spam [20], and Phishing-Fraud [31]. Phishing-Type is a collection of text data from 160 emails, each including the subject, text, and type of phishing email: fraud, false positives (legitimate emails), phishing, and commercial spam, with 40 emails in each type. Phishing-Spam comprises 5,685 email text messages, each associated with a binary label that indicates either spam or ham. Phishing-Fraud is an corpus that contains a total of 12,000 emails, each labeled as either fraudulent or legitimate. We randomly select n-shot samples for training, 20% for testing, and 20% for validation across all three datasets. The data statistics are summarized in Table 2.

Baselines We select some baselines that are most relevant to the scope and context of DRILL for comparison. These baselines include three widely recognized yet relatively small LMs: BERT [8], RoBERTa [27], and DistilBERT [32]. For each LM, we create two variants: one performing model fine-tuning, denoted as X-FT, and another freezing the model parameters but exclusively training the head, denoted as X-HD. Additionally, we include Distilling Step-by-Step (referred to as Distilling) [18] and WARP [15] which are specifically relevant to our method as they focus on rationale distillation and adversarial reprogramming, both of which utilize BERT as their backbone in our evaluation.

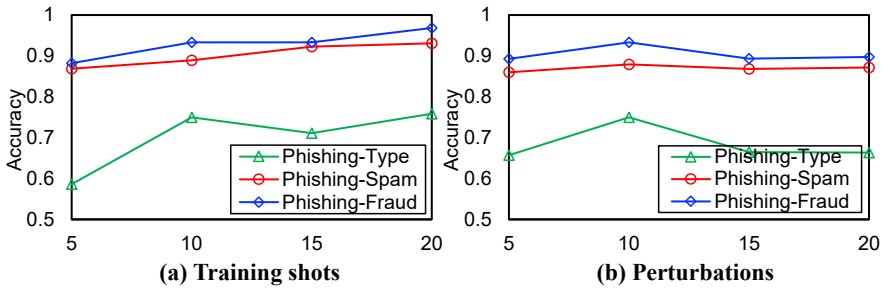

Fig. 2: Parameter evaluation in terms of phishing email detection accuracy: (a) impact of training shots and (b) impact of perturbation numbers.

Implementation Details We use GPT-3.5-Turbo as the LLM to generate rationales, and use BERT, RoBERTa, and DistilBERT as the small LMs for phishing email detection, respectively. The balance parameter for optimization is set as $\lambda = 0.5$. The most significant parameters that impact the performance of DRILL are the number of shots and the number of perturbations, which are evaluated in Section 4.2. All the results are reported using accuracy and F1-score.

4.2 Parameter Evaluation

We initiate our experimentation by assessing DRILL's phishing email detection performance with respect to different training shots n and perturbation numbers t. Specifically, we change these two parameters within the following ranges: $n \in \{5, 10, 15, 20\}$, and $t \in \{5, 10, 15, 20\}$, and report the results in Figure 2(a) and (b). It is worth noting that, to better facilitate the evaluations regarding the impact of individual parameters on model performance, when one parameter is tested, another one is fixed as 10.

- **Impact of training shots** As shown in Figure 2(a), when the number of training shots n increases, the detection accuracy tends to increase as well. Notably, the 10-shot scenario yields the first significant stabilization in performance, showcasing the largest increase in accuracy compared to other shot numbers. This suggests that a moderate amount of labeled data is sufficient to achieve robust performance without the diminishing returns observed with higher shot sizes.
- **Impact of perturbations** In Figure 2(b), we can observe that the performance peaks when $t = 10$ perturbations are appended to the inputs. This number of perturbations appears to be small yet optimal to harness the impact of the perturbations for enhancing model performance, which introduces variability while not substantially altering the original email representations. Beyond this point, performance declines consistently, implying that (1) larger perturbations cause excessive parameter updates, which are unsuitable for data-limited scenarios, and (2) extensive perturbations truncate the end of

Table 3: Comparison with Baselines (%): blue statistics denote the best results; all DRILL models are trained using $n = 10$ and $t = 10$.

Models	Phishing-Type		Phishing-Spam		Phishing-Fraud	
	ACC	F1	ACC	F1	ACC	F1
BERT-FT	31.25 ± 0.7	14.88 ± 0.7	70.99 ± 0.3	65.15 ± 0.4	54.91 ± 0.3	38.92 ± 0.3
RoBERTa-FT	40.62 ± 0.8	27.74 ± 0.8	72.36 ± 0.8	60.76 ± 0.8	54.93 ± 0.4	38.92 ± 0.4
DistilBERT-FT	31.25 ± 0.5	16.57 ± 0.9	87.40 ± 0.1	87.77 ± 0.1	90.61 ± 0.5	90.56 ± 0.5
BERT-HD	61.21 ± 0.5	61.37 ± 0.8	85.93 ± 0.3	85.77 ± 0.3	90.29 ± 0.3	88.31 ± 0.2
RoBERTa-HD	62.52 ± 0.7	60.88 ± 0.8	90.31 ± 0.7	84.87 ± 0.8	92.53 ± 0.4	92.50 ± 0.4
DistilBERT-HD	68.75 ± 0.5	68.08 ± 0.6	90.85 ± 0.1	90.78 ± 0.1	92.95 ± 0.2	91.22 ± 0.2
Distilling	65.00 ± 0.7	61.88 ± 0.5	89.53 ± 0.2	88.95 ± 0.2	92.60 ± 0.2	92.54 ± 0.2
WARP	66.50 ± 0.6	63.46 ± 0.8	91.11 ± 0.1	91.09 ± 0.1	92.96 ± 0.2	92.92 ± 0.1
DRILL-BERT	75.00 ± 1.8	74.87 ± 0.9	93.45 ± 0.1	93.43 ± 0.2	96.53 ± 0.1	96.52 ± 0.1
DRILL-RoBERTa	68.75 ± 0.7	68.40 ± 0.8	94.07 ± 0.2	94.04 ± 0.2	97.24 ± 0.1	97.20 ± 0.1
DRILL-DistilBERT	75.00 ± 0.6	74.85 ± 0.8	91.02 ± 0.2	90.89 ± 0.2	97.38 ± 0.1	97.31 ± 0.1

the email representation and larger perturbations overwrite substantial portions of the email, thereby compromising its semantic integrity. This decline highlights the delicate balance needed in perturbation size to maintain the efficacy of the original email content.

For these reasons, we choose to use $n = 10$ and $t = 10$ throughout the subsequent performance evaluations, as this combination provides a balance across model effectiveness, efficiency, and the integrity of the email representations.

4.3 Comparison with Baselines

To quantify the performance benefit of DRILL, we compare it against eight different baselines on three different datasets. For BERT, RoBERTa, and DistilBERT, we implement two variants for each of them: one that fine-tunes the entire LM (denoted as X-FT) and another that freezes the LM parameters while only training the head (denoted as X-HD). We report the comparative results in Table 3. A detailed comparison of these results reveals several key insights.

- **Baseline LM performance**. The baselines (BERT, RoBERTa, and DistilBERT) perform better when only the head is trained while their parameters are frozen. This suggests that fine-tuning the entire model is less effective with small training samples, likely due to overfitting. Training just the head allows the model to leverage the pretrained knowledge more effectively.
- **Advanced mechanisms** Among more advanced mechanisms, adversarial reprogramming outperforms rationale distillation, indicating that a smaller number of trainable parameters is more advantageous in data-limited scenarios. Adversarial reprogramming's ability to introduce subtle changes without overwhelming the model's pre-trained structure proves beneficial. Rationale distillation still requires large data to achieve better performance.
- **DRILL performance** DRILL shows significant performance improvements over BERT, RoBERTa, and DistilBERT, which also exceeds the accuracy of Distilling and WARP that leverage rationales and input token manipulation, respectively. This demonstrates the advantage of our dual-reasoning

scheme, which combines rationale distillation and trainable perturbations to effectively utilize limited data and enhance phishing email detection.

Overall, these findings highlight the effectiveness of DRILL in improving model performance in data-limited environments, showcasing its potential for real-world phishing email detection tasks.

We conduct further experimentation through an ablation study to verify the contribution of each aspect of DRILL to the model's learning process. In this experiment, we use BERT as a backbone LM, and investigate the impact of two key components in our model design: rationale distillation and trainable perturbations added to the inputs, and accordingly we construct three alternative models: (1) LM+Perturbations, (2) LM+Rationale, and (3) our complete model DRILL (note that, $n = 10$ is used across all models, and $t = 10$ is also used by models with perturbations).

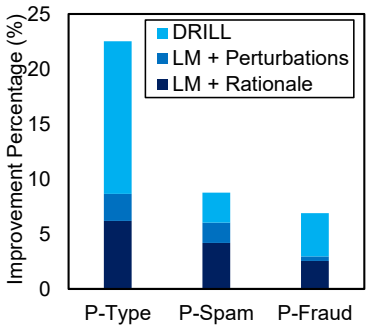

Fig. 3: Ablation Study: performance improvement percentages yielded by different components.

As illustrated in Figure 3, the inclusion of both components individually results in significant performance improvements compared to the BERT baseline. Specifically, incorporating rationale distillation into the model allows the small LM to benefit from the reasoning capabilities of the LLM, thus improving its understanding and classification of phishing emails. The rationale distillation ensures that the model learns from detailed explanations and task-specific knowledge provided by the LLM, bridging the reasoning gap that small LMs typically face. The addition of trainable perturbations to the inputs also contributes notably to the model's performance. This component introduces flexibility by allowing the model to adapt its behavior through interaction with the input tokens, enhancing the model's ability to capture intricate semantic patterns without altering the original parameters. The perturbations facilitate a more efficient adaptation to new tasks, even with limited data. When comparing the individual contributions, the perturbation component appears to be the larger contributor to performance gains. This is likely due to its direct impact on the model's adaptability and the reduction of trainable parameters, making it highly effective in data-limited scenarios.

Our proposed model, which integrates both rationale distillation and trainable perturbations, outperforms all other combinations. This demonstrates that each component is beneficial and necessary for the model's success. The dual-reasoning technique not only enhances the model's performance but also effectively mitigates the impact of data scarcity, showcasing the synergistic effect

Table 4: Case study: three representative models, including BERT-HD, Rationale Distillation, and DRILL-BERT are selected to identify if the given email is phishing or legitimate.

Email		**Rationale**
Dear Student We got your contact through your school database and I'm happy to inform you that we are currently running a student empowerment program to help students secure a part-time job with an attractive weekly wage. Kindly email back with your personal email address if interested in this job position.		The email exhibits phishing characteristics by requesting personal information and a personal email address which can be used for malicious purposes.
Model	**True Label**	**Predicted Label**
BERT-HD	Phishing	Legitimate
Rationale Distillation	Phishing	Legitimate
DRILL-BERT (*ours*)	Phishing	**Phishing**

of combining rationale distillation and perturbation-based adaptation. Overall, the ablation study confirms that the comprehensive design of DRILL is crucial for achieving superior performance in phishing email detection, particularly in environments with limited labeled data.

4.5 Case Study

To validate our claim that DRILL provides enhanced phishing email detection performance, we further perform a case study to showcase the different detection results offered by different models. Table 4 provides an email example along with its rationale, as well as the prediction results from BERT-HD, Rationale Distillation, and DRILL-BERT. In the given email example, the content is designed to deceive the recipient by requesting personal information under the guise of a student empowerment program. The rationale generated by our model identifies the phishing characteristics clearly, which highlights the specific elements that indicate phishing, such as the request for personal information and the use of a personal email address. In the case study, the prediction results from different models are as follows: (1) Despite having a sophisticated underlying model and the email exhibiting the clear phishing indicators, BERT-HD incorrectly labels the phishing email as legitimate; (2) similarly, the rationale distillation model, though equipped with the rationale that provides significant explanation to justify the phishing nature of the given email, also fails to generalize correctly due to the scarcity of training data, resulting in a misclassification as legitimate; (3) in contrast, our model DRILL-BERT successfully identifies the email as phishing by leveraging both the semantics of the email itself via perturbations and the rationale for enhancement.

This case study reaffirms the importance of dual-reasoning operations in DRILL. While traditional and rationale-based models struggle with generalization due to limited data, our approach successfully integrates perturbations and

rationales to adaptively and accurately identify phishing emails, demonstrating its superior performance and robustness in phishing email detection.

5 Ethical Statement

Similar to existing phishing email detection systems, the email data used may raise ethical concerns, particularly regarding privacy and the potential for false positives. However, we believe the merits of our work lie in reducing the need for extensive training email samples while enhancing detection performance, which not only mitigates user privacy concerns but also minimizes the risk of misclassifying legitimate emails.

6 Conclusion

In this paper, we present DRILL, a novel, simple yet effective approach that leverages dual-reasoning LLMs to enhance phishing email detection, particularly in data-limited scenarios. By incorporating rationales generated by LLMs and integrating trainable perturbations through small LMs, our method effectively bridges the gap between sophisticated reasoning capabilities and efficient training on limited datasets. Our extensive experimentation on three real-world email corpora demonstrate that DRILL outperforms traditional fine-tuning and head-training approaches for BERT, RoBERTa, and DistilBERT models, and delivers significant performance improvements when using our dual-reasoning technique compared to individual rationale distillation and adversarial reprogramming methods. Overall, DRILL offers a compelling solution for phishing email detection, demonstrating that dual-reasoning models can significantly enhance performance even with limited training data. This approach not only reduces the computational cost and labeled data requirements but also harnesses the power of pretrained models to extract intricate semantic patterns, making it a valuable tool in the fight against phishing and other potential cyber threats.

Acknowledgments

This work is partially supported by the NSF under grant CNS-2245968. The authors would also like to thank the reviewers for their valuable feedback.

References

1. Alhogail, A., Alsabih, A.: Applying machine learning and natural language processing to detect phishing email. Computers & Security **110**, 102414 (2021)
2. Altwaijry, N., Al-Turaiki, I., Alotaibi, R., Alakeel, F.: Advancing phishing email detection: A comparative study of deep learning models. Sensors **24**(7), 2077 (2024)

3. Bhardwaj, A., Sapra, V., Kumar, A., Kumar, N., Arthi, S.: Why is phishing still successful? Computer Fraud & Security **2020**(9), 15–19 (2020)
4. Brown, T., Mann, B., Ryder, N., Subbiah, M., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)
5. Cao, Y., Li, S., Liu, Y., Yan, Z., Dai, Y., Yu, P.S., Sun, L.: A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt. arXiv preprint arXiv:2303.04226 (2023)
6. Chen, L., Li, X., Wu, D.: Adversarially reprogramming pretrained neural networks for data-limited and cost-efficient malware detection. In: Proceedings of the 2022 SIAM International Conference on Data Mining (SDM). pp. 693–701. SIAM (2022)
7. Chen, Z., Mao, H., Wen, H., Han, H., Jin, W., Zhang, H., Liu, H., Tang, J.: Label-free node classification on graphs with large language models (llms). arXiv preprint arXiv:2310.04668 (2023)
8. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805 (2018)
9. Dhamija, R., Tygar, J.D., Hearst, M.: Why phishing works. In: Proceedings of the SIGCHI conference on Human Factors in computing systems. pp. 581–590 (2006)
10. Egozi, G., Verma, R.: Phishing email detection using robust nlp techniques. In: 2018 IEEE International Conference on Data Mining Workshops (ICDMW). pp. 7–12. IEEE (2018)
11. Elsayed, G.F., Goodfellow, I., Sohl-Dickstein, J.: Adversarial reprogramming of neural networks. arXiv preprint arXiv:1806.11146 (2018)
12. Fu, Y., Peng, H., Ou, L., Sabharwal, A., Khot, T.: Specializing smaller language models towards multi-step reasoning. In: International Conference on Machine Learning. pp. 10421–10430. PMLR (2023)
13. Greco, F., Desolda, G., Esposito, A., Carelli, A.: David versus goliath: Can machine learning detect llm-generated text? a case study in the detection of phishing emails
14. Hall, C.: Phishing email data by type. Kaggle, https://www.kaggle.com/datasets/charlottehall/phishing-email-data-by-type, version 1
15. Hambardzumyan, K., Khachatrian, H., May, J.: Warp: Word-level adversarial reprogramming. arXiv preprint arXiv:2101.00121 (2021)
16. Heiding, F., Schneier, B., Vishwanath, A., Bernstein, J., Park, P.S.: Devising and detecting phishing emails using large language models. IEEE Access (2024)
17. Ho, N., Schmid, L., Yun, S.Y.: Large language models are reasoning teachers. arXiv preprint arXiv:2212.10071 (2022)
18. Hsieh, C.Y., Li, C.L., Yeh, C.K., Nakhost, H., Fujii, Y., Ratner, A., Krishna, R., Lee, C.Y., Pfister, T.: Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In: ACL. p. 8003–8017 (2023)
19. Huang, J., Gu, S.S., Hou, L., Wu, Y., Wang, X., Yu, H., Han, J.: Large language models can self-improve. arXiv preprint arXiv:2210.11610 (2022)
20. Jackksoncsie: Spam email dataset. Kaggle, https://www.kaggle.com/datasets/jackksoncsie/spam-email-dataset, version 1
21. Kang, M., Lee, S., Baek, J., Kawaguchi, K., Hwang, S.J.: Knowledge-augmented reasoning distillation for small language models in knowledge-intensive tasks. NeurIPS **36** (2024)
22. Koide, T., Fukushi, N., Nakano, H., Chiba, D.: Chatspamdetector: Leveraging large language models for effective phishing email detection. arXiv preprint arXiv:2402.18093 (2024)
23. Kojima, T., Gu, S.S., Reid, M., Matsuo, Y., Iwasawa, Y.: Large language models are zero-shot reasoners. Advances in neural information processing systems **35**, 22199–22213 (2022)

24. Labonne, M., Moran, S.: Spam-t5: Benchmarking large language models for few-shot email spam detection. arXiv preprint arXiv:2304.01238 (2023)
25. Lester, B., Al-Rfou, R., Constant, N.: The power of scale for parameter-efficient prompt tuning. EMNLP (2021)
26. Li, S., Chen, J., Shen, Y., Chen, Z., Zhang, X., Li, Z., Wang, H., Qian, J., Peng, B., Mao, Y., et al.: Explanations from large language models make small reasoners better. arXiv preprint arXiv:2210.06726 (2022)
27. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
28. Magister, L.C., Mallinson, J., Adamek, J., Malmi, E., Severyn, A.: Teaching small language models to reason. arXiv preprint arXiv:2212.08410 (2022)
29. Muralidharan, T., Nissim, N.: Improving malicious email detection through novel designated deep-learning architectures utilizing entire email. Neural Networks **157**, 257–279 (2023)
30. Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. In: Proceedings of the Conference the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT). pp. 2227–2237 (2018)
31. Radev, D.: Fraud email dataset. Kaggle, https://www.kaggle.com/datasets/llabhishekll/fraud-email-dataset, version 1
32. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108 (2019)
33. Stojnic, T., Vatsalan, D., Arachchilage, N.A.: Phishing email strategies: understanding cybercriminals' strategies of crafting phishing emails. Security and privacy **4**(5), e165 (2021)
34. Timko, D., Castill, D.H., Rahman, M.L.: Unveiling human factors and message attributes in a smishing study. arXiv preprint arXiv:2311.06911 (2024)
35. Valecha, R., Mandaokar, P., Rao, H.R.: Phishing email detection using persuasion cues. TDSC **19**(2), 747–756 (2021)
36. Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., et al.: Emergent abilities of large language models. arXiv preprint arXiv:2206.07682 (2022)
37. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems **35**, 24824–24837 (2022)
38. Yasin, A., Abuhasan, A.: An intelligent classification model for phishing email detection. arXiv preprint arXiv:1608.02196 (2016)
39. Zavrak, S., Yilmaz, S.: Email spam detection using hierarchical attention hybrid deep learning method. Expert Systems with Applications **233**, 120977 (2023)
40. Zelikman, E., Wu, Y., Mu, J., Goodman, N.: Star: Bootstrapping reasoning with reasoning. NeurUPS **35**, 15476–15488 (2022)
41. Zhang, M., Sun, M., Wang, P., Fan, S., Mo, Y., Xu, X., Liu, H., Yang, C., Shi, C.: Graphtranslator: Aligning graph model to large language model for open-ended tasks. In: Proceedings of the ACM on Web Conference 2024. pp. 1003–1014 (2024)
42. Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H.: A comprehensive survey on transfer learning. Proceedings of the IEEE **109**(1), 43–76 (2020)