Do Spammers Dream of Electric Sheep? Characterizing the Prevalence of LLM-Generated Malicious Emails

Wei Hao
Columbia University
New York, United States
Barracuda Networks
Campbell, United States

Van Tran
University of Chicago
Chicago, United States

Vincent Rideout
Barracuda Networks
Campbell, United States

Zixi Wang
Columbia University
New York, United States
Barracuda Networks
Campbell, United States

AnMei Dasbach-Prisk[*]
University of California San Diego
San Diego, United States

M. H. Afifi
Barracuda Networks
Campbell, United States

Junfeng Yang
Columbia University
New York, United States

Ethan Katz-Bassett
Columbia University
New York, United States

Grant Ho
University of Chicago
Chicago, United States

Asaf Cidon
Columbia University
New York, United States
Barracuda Networks
Campbell, United States

Abstract

The rapid adoption of large language models (LLMs) has fueled speculation that cybercriminals may utilize LLMs to improve and automate their attacks. However, so far, the security community has had only anecdotal evidence of attackers using LLMs, lacking large-scale data on the extent of real-world malicious LLM usage.

In this joint work between academic researchers and Barracuda Networks, we present the first large-scale study measuring AI-generated attacks in-the-wild. In particular, we focus on the use of LLMs by attackers to craft the text of malicious emails by analyzing a corpus of hundreds of thousands of real-world malicious emails detected by Barracuda. The key challenge in this analysis is determining ground truth: we cannot know for certain whether an email is LLM or human-generated. To overcome this challenge, we observe that, prior to the launch of ChatGPT, email text was almost certainly not LLM-generated. Armed with this insight, we run three state-of-the-art LLM detection methods on our corpus and calibrate them against pre-ChatGPT emails, as well as against a diverse set of LLM-generated emails we create ourselves.

Since the launch of ChatGPT, all three detection methods indicate that attackers have steadily increased their use of LLMs to generate emails, especially for spam. Using our most precise AI-detection method, we conservatively estimate that *at least* ~51% of spam emails and ~14% of business email compromise attacks in our dataset are generated using LLMs, as of April 2025. Finally, analyzing the text of LLM-generated emails, we find evidence that attackers use LLMs to "polish" their emails and to generate multiple versions of the same email message.

CCS Concepts

• **Computing methodologies → Machine learning**; • **General and reference → Measurement**.

Keywords

LLM-generated Content Detection; Malicious Emails

ACM Reference Format:
Wei Hao, Van Tran, Vincent Rideout, Zixi Wang, AnMei Dasbach-Prisk, M. H. Afifi, Junfeng Yang, Ethan Katz-Bassett, Grant Ho, and Asaf Cidon. 2025. Do Spammers Dream of Electric Sheep? Characterizing the Prevalence of LLM-Generated Malicious Emails. In *Proceedings of the 2025 ACM Internet Measurement Conference (IMC '25), October 28–31, 2025, Madison, WI, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3730567.3732922

1 Introduction

Recent advances in LLMs have raised concerns about their potential misuse for illegal or unethical activities, such as spreading misinformation, manipulating social media, and scaling phishing campaigns. Among these, malicious emails pose a significant threat due to their potential for direct financial harm. According to the Internet Crime Complaint Center (IC3), phishing emails resulted in approximately $55 billion in losses between 2013 and 2023, with an increase of 9% in 2023 [35, 36]. Despite widespread speculation about how

[*] Work done at Columbia University.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

IMC '25, Madison, WI, USA
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1860-1/2025/10
https://doi.org/10.1145/3730567.3732922

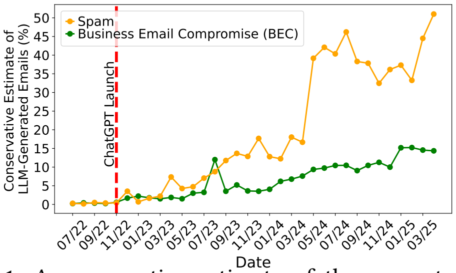


Figure 1: A conservative estimate of the percentage of LLM-generated messages across all malicious emails in our dataset. The red dotted line marks the launch of ChatGPT in 2022.

LLMs could be used to craft attack emails [9, 16–18, 21, 38], few studies have provided empirical data on how cyber criminals are using LLMs for this purpose and, more generally, how they are employing AI at scale for cyber attacks.

We present the first systematic analysis of LLM usage in malicious emails, using a large-scale real-world dataset of 481,558 malicious emails collected from both before and after the release of ChatGPT, which marked the beginning of widespread public access to LLMs. We employ state-of-the-art methods to detect LLM-generated content [4, 31, 39] to analyze the usage of LLMs chronologically across two categories of emails: spam and business email compromise (BEC). We validate the methods on a set of labeled ground-truth data that contains human-generated emails (sent before the launch of ChatGPT) and LLM-generated emails that we ourselves create. Among the three detection methods, one of them (fine-tuning an LLM for binary classification using the RoBERTa model [29]) exhibits near-zero false positives and false negatives on ground-truth data. Since it is extremely accurate in detecting human-generated emails, we assume that the emails it has detected as LLM-generated after the launch of ChatGPT represent a rough "floor" for the true LLM-generated rate. Using this method, we find that LLM-generated emails constitute at least 51% of spam emails in the last month of our dataset, while for BEC emails they comprise at least 14% of emails (Figure 1).

We explore the prevalent topics in each email category , highlighting the overlaps and distinctions between LLM-generated and human-generated emails. Additionally, we examine the intent behind LLM usage by comparing the characteristics of LLM-generated malicious emails with those crafted by humans, and we investigate the activities of top spammers employing LLMs. To summarize, we address the following key questions:

Q1: To what extent are LLM used to write malicious emails? (§4.3)

Q2: What are the prevalent topics in LLM-generated emails compared to human-generated ones? (§5.1)

Q3: How does the writing quality of LLM-generated malicious emails compare to human-generated ones? (§5.2)

Q4: What are some existing strategies for how attackers use LLMs in-the-wild? (§5.3)

2 Background and Related Work

This section provides background on LLM-generated content detection (§2.1), prior work measuring LLM-generated content (§2.2), and examining the potential use of LLMs for phishing (§2.3).

2.1 LLM-Generated Content Detection

We focus on three primary techniques with distinct features to identify and differentiate between human and LLM-generated text.

Fine-tuning an LLM for detection. The first technique we use to detect LLM-generated content is to fine-tune an existing LLM for binary classification on an input text set that consists of texts labeled as LLM or human-generated [13, 33, 39]. We adopt this technique using a fine-tuned RoBERTa model [29, 39].

Rewriting using a pre-trained LLM. The second method relies on the observation that, when LLMs are prompted to rewrite an input text, their output contains more changes when the input was written by a human than when it was LLM-generated (even when using a different LLM) [19, 31]. We adopt RAIDAR [31], a recent technique that prompts an LLM to rewrite input texts and uses the edit distance between the original and rewritten texts as a feature to train a logistic regression model for classifying human versus LLM-generated text.

Comparing the conditional probabilities of tokens. The third technique leverages the raw logit outputs from LLMs to compute the probability that the text is LLM-generated [4, 32, 40, 42]. We adopt Fast-DetectGPT [4], which assumes LLM-generated text outputs certain tokens at a higher probability conditioned on previous tokens. It calculates the conditional probability of the input tokens based on the previous ones and compares it to a threshold representing the conditional probability of token generation that would be typical of LLMs.

2.2 Measuring LLM-Generated Content

Recent work measuring LLM-generated text in-the-wild focuses on academic papers and reviews, analyzing trends before and after ChatGPT's release [26, 27]. Both works employ a word frequency-based method, which relies on having access to an accurate estimation of a constructed LLM-generated corpus during training. Researchers found that up to 17.5% of their post-GPT review corpus could have been modified by LLMs [26] and that up to 16.9% of the paper corpus could be LLM-generated or modified [27]. These works use distributional estimations over the entire corpus to compute how much is human-generated versus LLM-generated. They do not have a direct way to label individual text items (e.g., individual emails) as LLM or human-generated, and thus do not support the analysis conducted by our work (e.g., § 5).

Concurrent with our work, researchers at Mimecast used an in-house detector to analyze approximately 2,000 emails per month—both benign and malicious—over a span of about 40 months [44]. Their blog post reports similar findings, including a rise in LLM-generated content across all emails. Our study explores a variety of different state-of-the-art LLM-generated text detection methods on a larger dataset and sheds additional light on the differences between malicious LLM-generated and human-generated emails (§5.2). Taken together, these findings spanning different email datasets with different LLM-generated text detectors suggest significant attacker uptake in malicious LLM usage.

2.3 Generating Phishing Emails With LLMs

Several recent studies have investigated the use of LLMs for generating phishing emails [16–18, 21, 38]. Unlike our work, these studies

Taxonomy	Train 02/22-06/22	Test (Pre-GPT) 07/22-11/22	Test (Post-GPT) 12/22-04/25
Spam	14,646	11,751	212,748
BEC	11,616	18,450	212,347

Table 1: Number of emails used for training and testing in LLM-generated email detection for spam and BEC emails.

focus on hypothetical scenarios and do not measure real-world phishing data. Traditionally, phishing emails have been plagued by poor writing and grammatical errors [15, 22]. Crafting effective spear-phishing emails typically requires more effort, including collecting personal data, finely tailoring content, and mastering linguistic nuances [6, 21]. Recent work shows that LLMs can reduce these challenges by assisting with data collection [20, 24] and generating polished phishing emails with minimal effort [9].

To generate phishing emails, attackers can directly prompt these models [9, 21] or use LLMs to generate tailored malicious prompts, which attackers can then use to create phishing emails that resemble the communication style of well-known brands [38]. Attackers with access to legitimate or malicious email data can also train LLMs to produce realistic phishing emails that closely mimic authentic communications [12, 17, 18]. However, there is no consensus on whether LLM-generated phishing emails are more effective than those crafted by human experts [6, 21, 24].

3 Dataset

We now describe our dataset (§3.1), how it is preprocessed (§3.2), the ethical aspects of our study (§3.3), and the limitations (§3.4).

3.1 Email Categories

For our study, we collaborate with Barracuda Networks, a large security company, which we refer to as Barracuda in this paper. Our dataset consists of malicious emails sent to thousands of organizations that are Barracuda customers. These organizations span a diverse set of sectors including business services, construction, education, finance, government, healthcare, manufacturing, non-profit organizations, and retail. Collectively they employ millions of enterprise users. The emails were identified as malicious by two of Barracuda's commercial detection systems that use textual and URL-based features extracted from the email body. The systems achieve over 99% precision based on manual validation by Barracuda analysts. We consider two categories of malicious emails:

- **Spam:** unsolicited and untargeted emails sent to many recipients, often advertising unrealistic offers and enticing them to provide upfront fees or personal information.
- **Business Email Compromise (BEC):** targeted email attacks aimed at deceiving an individual within an organization to steal funds or sensitive information by impersonating a trusted figure (e.g., the recipient's manager or CEO).

Each category is detected by separately-trained detectors, and no emails belong to both categories.

3.2 Data Cleaning and Statistics

We selected emails written in English sent between February 2022 and April 2025, covering periods both pre- and post-launch of ChatGPT (Nov 30, 2022). We removed emails containing forwarded content to ensure each email contains a single message body. We processed the emails by extracting message text from the HTML body when applicable. We then applied Unicode normalization on the text and replaced all URLs with "[link]". Unless otherwise specified, we de-duplicated the emails based on their (Internet message ID [34], sender's email address, and email body). Finally, we filtered out emails that had fewer than 250 characters, since the text detectors are inaccurate on very short texts.

After all data cleaning steps, our final dataset consists of 239,145 emails for spam and 242,413 emails for BEC. Table 1 provides a breakdown of our dataset across time, including the specific training and test data splits (§4).

3.3 Ethics

This study involves a joint collaboration between academic researchers and Barracuda. The analysis utilized an email corpus from organizations that are active clients of Barracuda and gave permission to use their data for research purposes. All data and analysis took place on Barracuda's servers. Access to the data was restricted to designated researchers who are either active employees or contractors of Barracuda, as outlined in a data-sharing agreement between Barracuda and the researchers' institution, and enforced through robust access control measures.

3.4 Limitations

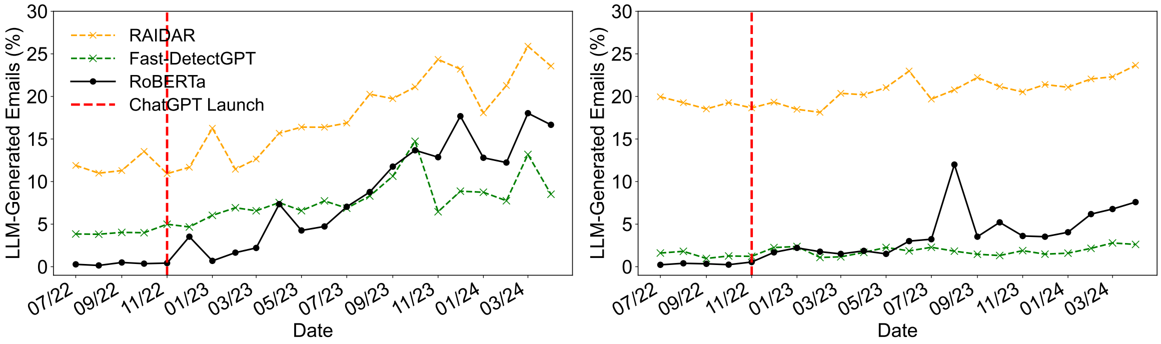
Although our dataset includes malicious emails collected from a diverse range of organizations, enhancing the generalizability of LLM-generated email detection [19], we rely on one organization's (Barracuda's) detection labels, which may produce a biased subset of malicious emails. Barracuda's detection systems evolve over time, making it difficult to disentangle the extent to which changes in the rate of LLM-generated emails we detect reflect changes in malicious LLM usage overall versus changes in what the systems flag as malicious. Nonetheless, our results clearly point to a substantial growth in the proportion of emails identified as LLM-generated, indicating widespread adoption of LLMs by attackers. Additionally, given the absence of ground-truth labels for LLM-generated emails, we constructed our training data of LLM-generated emails using proxy methods (see §4.1); it is possible that this approach produces a different distribution of emails than those generated by real-world attackers. These factors introduce potential bias into our detectors' training data, which may lead to an underestimation of the true prevalence or diversity of malicious LLM-generated emails.

4 Detecting LLM-Generated Emails

To identify LLM-generated emails, we used the three LLM detection methods mentioned in §2.1, RoBERTa, RAIDAR, and Fast-DetectGPT. We first describe how we trained the RoBERTa and RAIDAR detectors for our task, following the best practices set by prior work [13, 31, 33, 39] (§4.1). We then describe the calibration of our detection methods using data collected before the launch of ChatGPT (§4.2) and present our detection results (§4.3).

4.1 Training the Detectors

To train the models, we split each malicious email dataset into two sets: a training dataset consisting of five months of emails starting on February 2022 and a test dataset consisting of 34 months



(a) Detection for spam.

(b) Detection for BEC.

Figure 2: Percentage of emails detected as LLM-generated for Spam and BEC emails using RoBERTa, RAIDAR, and Fast-DetectGPT. The red dotted line marks the launch of ChatGPT in December 2022. The pre-GPT detection rate reflects the false positive rate of each method.

	RoBERTa	RAIDAR
Spam	0.0%/0.0%	9.6%/10.9%
BEC	0.1%/0.1%	15.3%/18.2%

Table 2: False positive rate/false negative rate of RoBERTa and RAIDAR on the validation datasets.

of emails starting on July 2022, as shown in Table 1. For model hyper-parameter training, we further randomly split each training dataset and use 80% of data for training and 20% of data for validation (parameter tuning). Both RoBERTa and RAIDAR need a labeled training dataset. Since the training dataset only contains emails before the launch of ChatGPT, we treat all these emails as human-generated (non-LLM) and expand this training data with LLM-generated emails that we generate from the human-generated ones. Generating LLM-based malicious emails from human-written examples may not fully reflect all real-world LLM-generated attack emails, which could lead to inaccurate estimations. However, there is currently no ground truth on how attackers generate such content, and our approach provides a best-effort approximation of this process. Due to Barracuda's policies, we cannot pass the emails to a third-party commercial LLM, so we use Mistral-7B-Instruct-v0.2 [23] hosted locally to create LLM-generated emails, using the default temperature of 1. For this generation, we prompt the model to rewrite an existing human-generated malicious email. Appendix A.3 contains the specific prompts we used.

Recall from §2 that, in addition to a training dataset of LLM and human-generated text, RAIDAR needs to rewrite each text input (to compute the edit distance for classification). To rewrite each email for RAIDAR's training and classification, we use a different open-source model, Llama-2-7b-chat [3], to capture the real-world scenario in which the generation model and rewriting model may not be the same. The system prompt and rewriting prompts we used for RAIDAR can be found in Appendix A.3. We use a generation temperature of 0 for rewriting to enhance determinism.

Given these training datasets, we train separate RoBERTa and RAIDAR detectors for each category of malicious emails, continuing training until the models converge on their validation datasets. We stop training when the model accuracy remains consistent for three consecutive epochs. The validation results are shown in Table 2. When running RAIDAR, we limit each email to the first 2,000 characters to prevent out-of-memory issues though it constrains RAIDAR's ability to detect emails where LLMs are only used after the text cutoff. For Fast-DetectGPT, which does not require training, we use the open source version [1], due to its reported ability to robustly detect LLM-generated content from many different models, including GPT-4, across diverse text domains [4].

4.2 False Positive Rates

A very important metric for any detector and for our analysis is the false positive rate, which is the number of human-generated emails the detector incorrectly classifies as LLM-generated. If a detector achieves a low false positive rate on the five months of ground truth data that predate ChatGPT, we can trust that the detections it makes during the post-ChatGPT phase likely represent a rough "lower bound" on the percentage of LLM-generated emails.

Figure 2 shows that the false positive rates of the three detectors remain relatively flat during the entire pre-ChatGPT period across both spam and BEC. This is important, because it suggests that the false positive rate will continue to remain flat post-ChatGPT. Interestingly, the false positive rates vary significantly across approaches. RoBERTa yields by far the lowest false positive rates (0.3% on spam and 0.4% on BEC), followed by Fast-DetectGPT (4.3% and 1.4%). RAIDAR exhibits high false positive rates (11.7% and 19.1%). There is no reason to believe that the characteristics of genuine human-generated emails (and consequently the false positive rate) would change drastically after ChatGPT's launch. We therefore conclude that RoBERTa is a very precise method of detecting LLM-generated emails. It may not detect all or most of the LLM-generated emails (indeed, some prior work claims that binary classification models do not generalize well and may miss content generated by models that are different than the ones used in training [37, 41]), but it is very useful for our study, as it suffers from a *near-zero false positive rate*. Therefore, it can serve as an effective *lower bound* for the number of LLM-generated emails in our dataset. Armed with this insight, we now analyze the prevalence and growth of LLM-generated malicious emails following ChatGPT's launch.

4.3 Detection Results

Figure 2 shows results from July 2022 through April 2024 for all three detection methods. If we use RoBERTa, the most conservative method, as our detector, the percentage of spam that is LLM-generated in April 2024 is at least 16.2%, and the percentage of BEC that is LLM-generated is at least 7.6%. To conservatively assess if LLM usage has further increased over time, we apply RoBERTa to more recent data and find continued growth in LLM use, with 51% of spam emails and 14.4% of BEC emails flagged as LLM-generated in April 2025 (Figure 1). As mentioned in §4.2, since we suspect RoBERTa may miss a non-negligible percent of LLM-generated emails, the true proportion of LLM-generated emails is likely higher.

All three detectors yield qualitatively similar results: Both types of email show a relatively steady increase in LLM use over time, but the rate of increase for spam is much faster. These trends are evident even with the "noisier" detectors (RAIDAR and Fast-DetectGPT). Since spam is less targeted than BEC and requires less expertise in crafting the emails, the lower increase in usage of LLM for BEC could suggest that LLMs are not used as frequently in spear phishing as compared to spam. The rate of emails generated by LLM spikes for BEC in August 2023 and for spam in May 2024, which we suspect was driven by a combination of factors, including active spam/BEC campaigns, the launch of GPT-4o [43] in May 2024 leading to changes in attacker behavior, and updates to Barracuda's detection systems. We show examples of BEC and spam emails that were classified as LLM-generated in Figure 3 and in Appendix A.2.

To evaluate whether the increase in LLM usage is statistically significant, we conducted a Kolmogorov-Smirnov (K-S) test comparing the distributions of RoBERTa's predicted probabilities (for if a message is LLM-generated) on the emails before and after the launch of ChatGPT. The results indicate that the two distributions are statistically significantly different for both spam and BEC ($P < 0.001$).

5 Characterizing LLM-generated Malicious Emails

In this section, we analyze the topics and linguistic features of LLM-generated versus human-generated malicious emails to understand potential differences in these emails. From this broader analysis, we also present one case study that illustrates a use case of LLMs for attackers. Due to data access and compute constraints, we focus solely on emails in the post-GPT period up until April 2024. For our analysis, we label an email as LLM-generated if at least two of the three detectors label it as such; otherwise we label it as human-generated. By requiring the agreement of two detectors, we seek to minimize false positives and false negatives to best compare the contents of LLM-generated versus human-generated emails. This approach flags 2,812 spam emails and 1,940 BEC emails as LLM-generated. Figure 4 in Appendix A.1 shows that 88% and 87% of these emails are detected by RoBERTa, our most conservative approach with extremely low false positives. To reduce computation costs, we randomly downsampled the human-generated emails to have the same number as LLM-generated emails for our analysis.

5.1 Topic Modeling

We apply Latent Dirichlet Allocation (LDA) [7] to identify whether the message topics differ between LLM and human-generated emails. We run a separate LDA topic model for each email category (spam and BEC) and each set of human and LLM-generated emails (four models in total). We perform standard NLP cleaning steps (tokenization, stopwords removal, and lemmatization) and a standard hyperparameter grid-search on learning decay and the number of topics following prior work [2].

For BEC, both LLM and human-generated emails share the same most popular topics, including asking victims to update payroll information (55–55.9% of emails), or buy gift cards (4.6–7.8%), or pretending to be stuck in a meeting and asking for alternative text communication for further task assignment (27.9–32.3%).

For spam, we observe differences: LLM-generated emails primarily focus on promotional content about various products (82.7% of emails), while human-generated emails have an equal focus on promotional content as well as scams that ask victims to claim a fund or reward (40.9% and 42.2% respectively). Only 10.7% of LLM-generated emails contain these scams. Figure 3 shows some example spam and BEC emails, and Appendix A.2 provides additional examples with the top-10 salient terms identified by LDA.

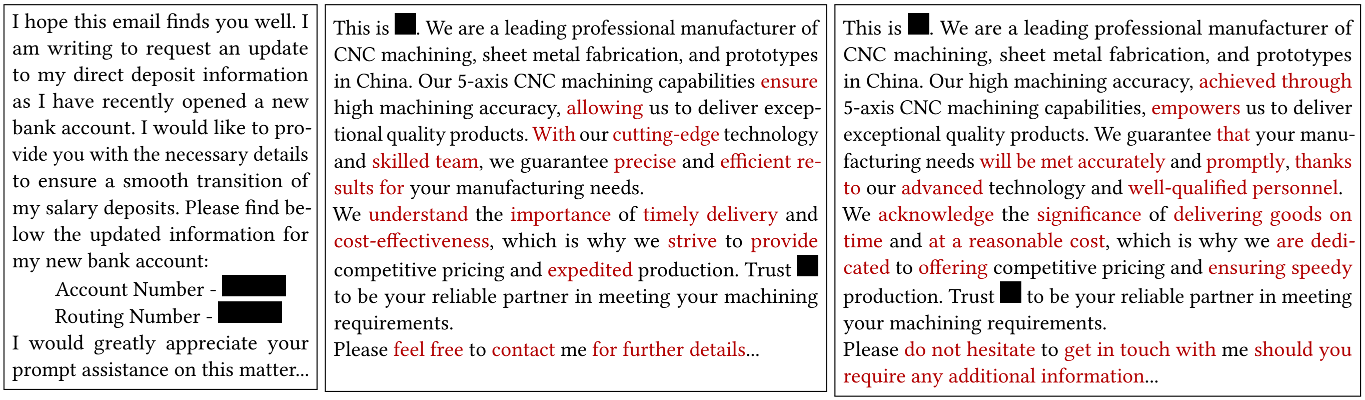
5.2 Linguistic Analysis

Security researchers have speculated that attackers may employ LLMs to generate more persuasive and effective email attacks [5, 10]. We now analyze the linguistic characteristics of the emails to assess whether significant differences do exist between human and LLM-generated emails, and whether these differences align with making the emails more effective. Specifically, we analyze features that capture aspects of writing quality (formality, sophistication, and grammatical errors) and tone (urgency) that could potentially impact the efficacy of phishing. None of the detectors we used to identify LLM-generated emails (§2.1) explicitly target or incorporate these linguistic features.

- **Formality**, scored from 1 to 5, describes whether the tone of an email is casual or formal.
- **Sophistication (Flesch reading-ease score [14])** rates English text readability from 0 to 100, with higher scores indicating easier comprehension.
- **Grammar-error** estimates the number of grammar errors [25], normalized between 0 and 1.
- **Urgency**, scored from 1 to 5, describes whether the tone of an email pressures the user into performing some kind of imminent action, such as clicking a link.

To score both formality and urgency, we use an LLM-based evaluation approach with a Llama-3.1-8B-Instruct model [3] as the evaluator. We prompt the model to score emails using instructions based on standard NLP work, such as G-Eval [30] and LLM-Eval [28]. Higher scores mean higher formality or urgency respectively. Our prompt defines each evaluation metric along with descriptions and a structured output schema (Figure 10 in Appendix A.3).

To check the accuracy of the LLM evaluations, two researchers independently scored a sample of 10 emails and compared their scores against each other and the LLMs. For urgency, the two researchers' Cohen Kappa score [11] was 0.63, and each researcher's agreement scores against the LLM's were 0.5 and 0.6. For formality, the alignment between LLM-generated and the human raters was positive but lower (0.19 and 0.67 respectively), and the score between the two human raters was 0.61. While the 1–5 scale allows finer distinctions, it also increases the chance of disagreement, resulting in lower agreement scores. Despite this, the agreement between LLMs and human raters—particularly for urgency—is comparable to the agreement between human raters themselves, suggesting LLMs



I hope this email finds you well. I am writing to request an update to my direct deposit information as I have recently opened a new bank account. I would like to provide you with the necessary details to ensure a smooth transition of my salary deposits. Please find below the updated information for my new bank account:

Account Number - [redacted]
Routing Number - [redacted]

I would greatly appreciate your prompt assistance on this matter...

This is [redacted]. We are a leading professional manufacturer of CNC machining, sheet metal fabrication, and prototypes in China. Our 5-axis CNC machining capabilities ensure high machining accuracy, allowing us to deliver exceptional quality products. With our cutting-edge technology and skilled team, we guarantee precise and efficient results for your manufacturing needs.
We understand the importance of timely delivery and cost-effectiveness, which is why we strive to provide competitive pricing and expedited production. Trust [redacted] to be your reliable partner in meeting your machining requirements.
Please feel free to contact me for further details...

This is [redacted]. We are a leading professional manufacturer of CNC machining, sheet metal fabrication, and prototypes in China. Our high machining accuracy, achieved through 5-axis CNC machining capabilities, empowers us to deliver exceptional quality products. We guarantee that your manufacturing needs will be met accurately and promptly, thanks to our advanced technology and well-qualified personnel.
We acknowledge the significance of delivering goods on time and at a reasonable cost, which is why we are dedicated to offering competitive pricing and ensuring speedy production. Trust [redacted] to be your reliable partner in meeting your machining requirements.
Please do not hesitate to get in touch with me should you require any additional information...

Figure 3: Examples of emails ("..." indicates omitted text, for brevity) detected as LLM-generated where we censor personalized details with ■. The first one is a BEC email; the second and third are spam emails. The spam emails seem to be reworded variants, with deltas colored in red.

Feature	human-generated		LLM-generated		P-Value	
	BEC	Spam	BEC	Spam	BEC	Spam
Formality (1-5)	3.6	3.3	3.9	4.0	< 0.001	< 0.001
Urgency (1-5)	3.0	2.1	3.0	1.5	0.32	< 0.001
Sophistication (0-100)	61.7	56.9	60.3	46.3	< 0.001	< 0.001
Grammar-error (0-1)	0.03	0.05	0.02	0.03	< 0.001	< 0.001

Table 3: The mean values of the linguistic features and the p-values of the KS-test comparing LLM versus human-generated for BEC and spam emails (§5.2). P-value<0.05 indicates statistical significance.

can reliably evaluate urgency, and to a lesser extent, formality, in malicious emails. When using a binary scale (<3 vs. ≥ 3), the Kappa score between LLMs and human raters reaches 1.0 for urgency and 0.9 for formality, indicating strong alignment.

To assess whether human and LLM-generated emails had significant differences across these features, we ran a KS-test on the scores for each of the four features. The results indicate statistically-significant differences across all linguistic features (Table 3) for spam and all features except for urgency for BEC. Specifically, LLM-generated emails were found to be more grammatically correct and formal, and they used more sophisticated language compared to human-generated emails. Thus, our measurements partially validate the hypothesis that malicious LLM-generated emails contain more sophisticated language. However, in our dataset, it does not appear that attackers are using LLMs to imbue their emails with specific persuasive elements like urgency, and LLM-generated spams were found even more neutral than the human-generated ones.

5.3 Spam: LLM Usage Case Study

When examining samples of spam emails during topic modeling and linguistic analysis, we observed an interesting trend. Many groups of LLM-generated spams appeared to be reworded variants of each other, where such rewording might aim to bypass spam filters by varying the word choice (presumably to avoid a volume-based filter that looks for identical emails being sent at a high volume, or perhaps to trick a filter that looks for specific combinations of words). Figure 3 shows two side-by-side examples. The emails have the same structure and content but slightly different wording.

To study this phenomenon further, we identify the top-100 malicious senders after the launch of ChatGPT by volume (after deduplicating emails by their Internet message ID and cleaned message content). These top-100 senders sent 25,929 unique spam messages. We then clustered the post-GPT emails from these top spammers using the MinHash locality-sensitive hashing [8], which clusters the text (email messages) by approximating the Jaccard similarity between the sets of words in each email. The five largest clusters contain between 668 to 1263 emails. Within these clusters, the percent of emails labeled as LLM-generated by a majority of our detectors is: 78.9%, 52.1%, 8.4%, 8.4% and 6.6%, with two of these containing significantly higher percentages of LLM-generated emails than the average percent of LLM-generated emails across all post-ChatGPT spam (7.8%). Next, we randomly sampled and analyzed five LLM-generated emails from the two clusters with extremely high percentages of LLM-generated emails (78.9% and 52.1%). From the first cluster, all five samples are rewritten versions of the same message; and for the second cluster, three out of the five samples are rewritten versions of each other. Appendix A.4 displays the messages from these clusters of rewritten emails. These results suggest that at least some attackers appear to use LLMs to generate many variants of the same message, potentially to evade detection.

6 Conclusions

We provide the first empirical investigation of how cyber criminals are employing AI at scale. We characterize a large-scale corpus of real-world malicious emails to answer one central question: to what extent are attackers using LLMs to generate email content in-the-wild? While BEC attacks exhibited a more modest amount of detected LLM usage, the usage of LLMs for generating spam has increased significantly, to the point where LLM-generated emails account for the majority of spam emails as of April 2025. Our work also sheds light on some potential ways attackers use LLMs, such as generating many versions of the same malicious message. Several open questions remain, including whether the malicious content produced by LLMs leads to a concrete increase in harm, e.g., by fooling more users or by evading current detectors.

Acknowledgments

This work was supported by Google, NSF award 2403431, the Columbia SEAS/EVPR Stimulus Award, and the Columbia SEAS-KFAI Generative AI Research Award. We thank Nick Feamster for supporting Van Tran's participation in this project. We thank our shepherd Alan Mislove and the anonymous reviewers for their insightful feedback that helped improve our work.

References

[1] 2024. Fast-DetectGPT. https://github.com/baoguangsheng/fast-detect-gpt

[2] Lin Ai, Sameer Gupta, Shreya Oak, Zheng Hui, Zizhou Liu, and Julia Hirschberg. 2024. TweetIntent@ Crisis: A Dataset Revealing Narratives of Both Sides in the Russia-Ukraine Crisis. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 18. 1872–1887.

[3] AI@Meta. 2024. Llama 3 Model Card. (2024). https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md

[4] Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. 2024. Fast-DetectGPT: Efficient Zero-Shot Detection of Machine-Generated Text via Conditional Probability Curvature. In *The Twelfth International Conference on Learning Representations*.

[5] Clark Barrett, Brad Boyd, Elie Bursztein, Nicholas Carlini, Brad Chen, Jihye Choi, Amrita Roy Chowdhury, Mihai Christodorescu, Anupam Datta, Soheil Feizi, Kathleen Fisher, Tatsunori Hashimoto, Dan Hendrycks, Somesh Jha, Daniel Kang, Florian Kerschbaum, Eric Mitchell, John Mitchell, Zulfikar Ramzan, Khawaja Shams, Dawn Song, Ankur Taly, and Diyi Yang. 2023. Identifying and Mitigating the Security Risks of Generative AI. *Foundations and Trends® in Privacy and Security* 6, 1 (2023), 1–52. doi:10.1561/3300000041

[6] Mazal Bethany, Athanasios Galiopoulos, Emet Bethany, Mohammad Bahrami Karkevandi, Nishant Vishwamitra, and Peyman Najafirad. 2024. Large Language Model Lateral Spear Phishing: A Comparative Study in Large-Scale Organizational Settings. *arXiv preprint arXiv:2401.09727* (2024).

[7] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3, Jan (2003), 993–1022.

[8] Andrei Z Broder. 1997. On the Resemblance and Containment of Documents. In *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No. 97TB100171)*. IEEE, 21–29.

[9] Stephanie Carruthers. 2024. AI vs. Human Deceit: Unravelling the New Age of Phishing Tactics. https://securityintelligence.com/x-force/ai-vs-human-deceit-unravelling-new-age-phishing-tactics/

[10] Google Cloud. 2024. Google Cybersecurity Forecast. https://cloud.google.com/security/resources/cybersecurity-forecast

[11] Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement* 20 (1960), 37 – 46. https://api.semanticscholar.org/CorpusID:15926286

[12] Avisha Das and Rakesh Verma. 2019. Automated email generation for targeted attacks using natural language. *arXiv preprint arXiv:1908.06893* (2019).

[13] Tiziano Fagni, Fabrizio Falchi, Margherita Gambini, Antonio Martella, and Maurizio Tesconi. 2021. TweepFake: About Detecting Deepfake Tweets. *PLOS One* 16, 5 (2021), e0251415.

[14] Rudolf Franz Flesch. 1948. A New Readability Yardstick. *Journal of Applied Psychology* 32 3 (1948), 221–33. https://api.semanticscholar.org/CorpusID:39344661

[15] Canadian Centre for Cyber Security. 2022. Don't take the bait: Recognize and avoid phishing attacks - ITSAP.00.101. https://www.cyber.gc.ca/en/guidance/dont-take-bait-recognize-and-avoid-phishing-attacks. Accessed: 2024-11-08.

[16] Alberto Giaretta and Nicola Dragoni. 2020. Community Targeted Phishing: A Middle Ground Between Massive and Spear Phishing Through Natural Language Generation. In *Proceedings of 6th International Conference in Software Engineering for Defence Applications: SEDA 2018 6*. Springer, 86–93.

[17] Shih-Wei Guo, Tzu-Chi Chen, Hui-Juan Wang, Fang-Yie Leu, and Yao-Chung Fan. 2022. Generating Personalized Phishing Emails for Social Engineering Training Based on Neural Language Models. In *International Conference on Broadband and Wireless Computing, Communication and Applications*. Springer, 270–281.

[18] Shih-Wei Guo and Yao-Chung Fan. 2024. X-Phishing-Writer: A Framework for Cross-Lingual Phishing Email Generation. *ACM Transactions on Asian and Low-Resource Language Information Processing* (2024).

[19] Wei Hao, Ran Li, Weiliang Zhao, Junfeng Yang, and Chengzhi Mao. 2024. Learning to rewrite: Generalized llm-generated text detection. *arXiv preprint arXiv:2408.04237* (2024).

[20] Julian Hazell. 2023. Spear Phishing With Large Language Models. *arXiv preprint arXiv:2305.06972* (2023).

[21] Fredrik Heiding, Bruce Schneier, Arun Vishwanath, Jeremy Bernstein, and Peter S Park. 2023. Devising and Detecting Phishing: Large Language Models vs. Smaller Human Models. *arXiv preprint arXiv:2308.12287* (2023).

[22] Cristian Iuga, Jason RC Nurse, and Arnau Erola. 2016. Baiting the hook: factors impacting susceptibility to phishing attacks. *Human-centric Computing and Information Sciences* 6 (2016), 1–20.

[23] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825* (2023).

[24] Hanna Kim, Minkyoo Song, Seung Ho Na, Seungwon Shin, and Kimin Lee. 2024. When LLMs Go Online: The Emerging Threat of Web-Enabled LLMs. *arXiv preprint arXiv:2410.14569* (2024).

[25] Languagetool-Org. 2024. Languagetool: Style and grammar checker for 25+ languages. https://github.com/languagetool-org/languagetool

[26] Weixin Liang, Zachary Izzo, Yaohui Zhang, Haley Lepp, Hancheng Cao, Xuandong Zhao, Lingjiao Chen, Haotian Ye, Sheng Liu, Zhi Huang, Daniel Mcfarland, and James Y. Zou. 2024. Monitoring AI-Modified Content at Scale: A Case Study on the Impact of ChatGPT on AI Conference Peer Reviews. In *Proceedings of the 41st International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 235)*, Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (Eds.). PMLR, 29575–29620. https://proceedings.mlr.press/v235/liang24b.html

[27] Weixin Liang, Yaohui Zhang, Zhengxuan Wu, Haley Lepp, Wenlong Ji, Xuandong Zhao, Hancheng Cao, Sheng Liu, Siyu He, Zhi Huang, Diyi Yang, Christopher Potts, Christopher D Manning, and James Y. Zou. 2024. Mapping the Increasing Use of LLMs in Scientific Papers. In *First Conference on Language Modeling*. https://openreview.net/forum?id=YX7QnhxESU

[28] Yen-Ting Lin and Yun-Nung Chen. 2023. LLM-Eval: Unified Multi-Dimensional Automatic Evaluation for Open-Domain Conversations With Large Language Models. *arXiv preprint arXiv:2305.13711* (2023).

[29] Yinhan Liu. 2019. Roberta: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692* 364 (2019).

[30] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-Eval: NLG Evaluation Using GPT-4 With Better Human Alignment. *arXiv preprint arXiv:2303.16634* (2023).

[31] Chengzhi Mao, Carl Vondrick, Hao Wang, and Junfeng Yang. 2024. RAIDAR: generative AI detection via rewriting. In *The Twelfth International Conference on Learning Representations*.

[32] Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. 2023. DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature. In *International Conference on Machine Learning*. PMLR, 24950–24962.

[33] Sandra Mitrović, Davide Andreoletti, and Omran Ayoub. 2023. Chatgpt or human? detect and explain. explaining decisions of machine learning model for detecting short chatgpt-generated text.

[34] o365devx. [n. d.]. InternetMessageId — learn.microsoft.com. https://learn.microsoft.com/en-us/exchange/client-developer/web-service-reference/internetmessageid. [Accessed 22-11-2024].

[35] Federal Bureau of Investigation. 2023. Business Email Compromise: The $50 Billion Scam. https://www.ic3.gov/PSA/2023/PSA230609

[36] Federal Bureau of Investigation. 2024. Business Email Compromise: The $55 Billion Scam. https://www.ic3.gov/PSA/2024/PSA240911

[37] Jiameng Pu, Zain Sarwar, Sifat Muhammad Abdullah, Abdullah Rehman, Yoonjin Kim, Parantapa Bhattacharya, Mobin Javed, and Bimal Viswanath. 2023. Deepfake text detection: Limitations and opportunities. In *2023 IEEE symposium on security and privacy (SP)*. IEEE, 1613–1630.

[38] Sayak Saha Roy, Poojitha Thota, Krishna Vamsi Naragam, and Shirin Nilizadeh. 2024. From Chatbots to Phishbots?: Phishing Scam Generation in Commercial Large Language Models. In *2024 IEEE Symposium on Security and Privacy (SP)*. 36–54. doi:10.1109/SP54263.2024.00182

[39] Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, et al. 2019. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203* (2019).

[40] Jinyan Su, Terry Yue Zhuo, Di Wang, and Preslav Nakov. 2023. DetectLLM: Leveraging Log Rank Information for Zero-Shot Detection of Machine-Generated Text. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

[41] Adaku Uchendu, Thai Le, Kai Shu, and Dongwon Lee. 2020. Authorship Attribution for Neural Text Generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 8384–8395. doi:10.18653/v1/2020.emnlp-main.673

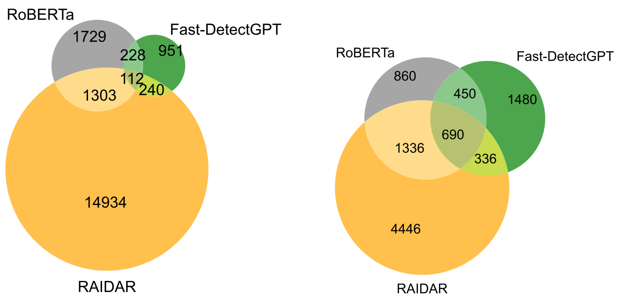
[42] Vivek Verma, Eve Fleisig, Nicholas Tomlin, and Dan Klein. 2024. Ghostbuster: Detecting Text Ghostwritten by Large Language Models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 1702–1717.

[43] Wikipedia contributors. 2025. GPT-4o — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/wiki/GPT-4o. https://en.wikipedia.org/wiki/GPT-4o Accessed on 2025-05-08.

[44] Andrew Williams. 2025. *Mimecast Threat Intelligence: How ChatGPT Upended Email.* https://www.mimecast.com/blog/how-chatgpt-upended-email/ Accessed: 2025-05-05.

A APPENDIX

A.1 Majority Voting Result among RoBERTa, RAIDAR and Fast-DetectGPT



BEC

Spam

Figure 4: Venn diagrams depicting majority voting for our three detectors for BEC (left) and spam (right) in §5. Each circle depicts the emails detected as LLM-generated by the corresponding detector.

Our analysis in §5 only considers emails detected as LLM-generated by at least two of our three detectors. Figure 4 depicts the agreement among the techniques. It is notable that 87% of flagged BEC emails and 88% of flagged spam emails are detected by RoBERTa, our most conservative approach with extremely low false positives. The less conservative approaches only contribute the 12-13% of emails that they *both* label as LLM-generated while RoBERTa does not.

A.2 Topic Modeling and Email Examples for BEC and Spam

This section includes more detailed Latent Dirichlet Allocation (LDA) outputs for the topic modeling in §5.1 and some example emails corresponding to these topics, shown in Figures 5 – 8.

Tables 4 and 5 show the top-ten terms identified by LDA for BEC and spam emails across human and LLM-generated emails. Each row corresponds to one LDA-discovered topic, though some topics may overlap or share similar terms. We performed a standard hyper-parameter grid search for our LDA model [2], on learning decay (0.5-0.9) and the number of topics (2-16), with topic coherence as the evaluation metric following prior work [2].

For BEC, both LLM and human-generated emails share the same most popular topics, including asking victims to update payroll information, ('direct deposit', 'payroll' and 'bank': 55% of LLM-generated and 55.9% of human-generated emails contain these terms), buying gift cards ('gift' and 'card': 4.6% of LLM-generated and 7.8% of human-generated emails), and pretending to be stuck in a meeting and asking for alternative text communication for further task assignment ('meeting', 'mobile', 'cell', 'phone' and 'task': 32.3% of LLM-generated and 27.9% of human-generated emails contain these terms). Figures 5 and 6 show examples of these emails.

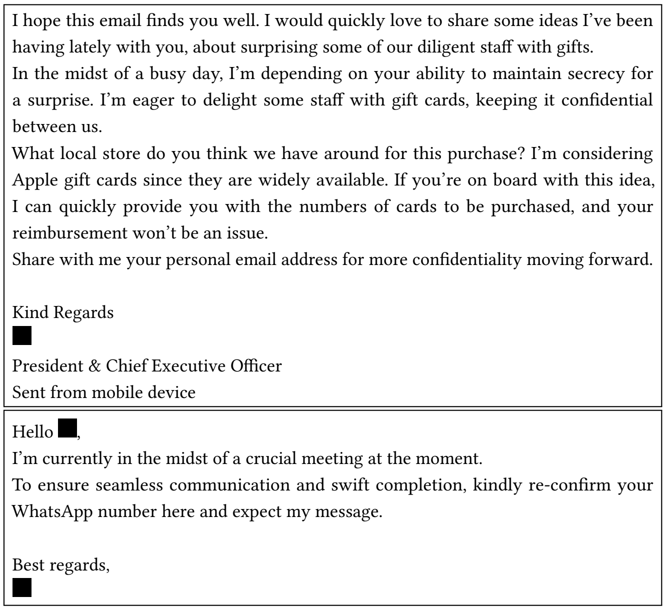
For spam, we observe differences in the topics for LLM-generated versus human-generated emails. LLM-generated emails primarily focus on promotional content about various products ('manufacturer', 'manufacturing', 'design', 'supply', 'solution': 82.7% of LLM-generated emails contain these thematic terms, compared to only 40.9% of human-generated emails), while human-generated emails focus more on scams that ask victims to claim a fund or reward ('fund', 'bank', 'million', 'payment': 42.2% of human-generated emails contain these scam-related terms, compared to only 10.7% of LLM-generated emails). Figure 7 and 8 show examples of these emails.

human-generated	**LLM-generated**
account, bank, new, deposit, direct, change, information, detail, like, update	information, deposit, direct, account, change, bank, update, new, detail, request
information, direct, deposit, change, detail, payroll, pay, update, banking, new	number, phone, soon, need, possible, hope, find, currently, convenience, sincerely
number, task, meeting, kindly, response, phone, cell, executive, text, learn	kindly, need, number, meeting, best, message, cell, task, currently, text
gift, card, let, know, surprise, sent, purchase, mobile, device, today	know, sent, gift, today, best, mobile, card, device, appreciate, let,

Table 4: Top 10 salient terms per topic generated by LDA for BEC emails. Each row corresponds to one LDA-discovered topic, though some topics may overlap or share similar terms. Number of topics is picked by the grid search result.

human-generated	**LLM-generated**
product, mold, quality, company, best, manufacturer, china, high, price, part	bag, paper, business, packaging, website, manufacturer, best, interested, contact, team
fund, bank, united, business, contact, address, million, payment, dollar, state	mold, manufacturing, cnc, casting, machining, offer, team, including, range, design
	led, cost, renerge, development, driver, supply, power, custom, manufacturer, procurement
	best, customer, business, forward, manufacturer, industry, china, solution, contact, meet

Table 5: Top 10 salient terms per topic generated by LDA for spam emails. Each row corresponds to one LDA-discovered topic, though some topics may overlap or share similar terms.Number of topics is picked by the grid search result.



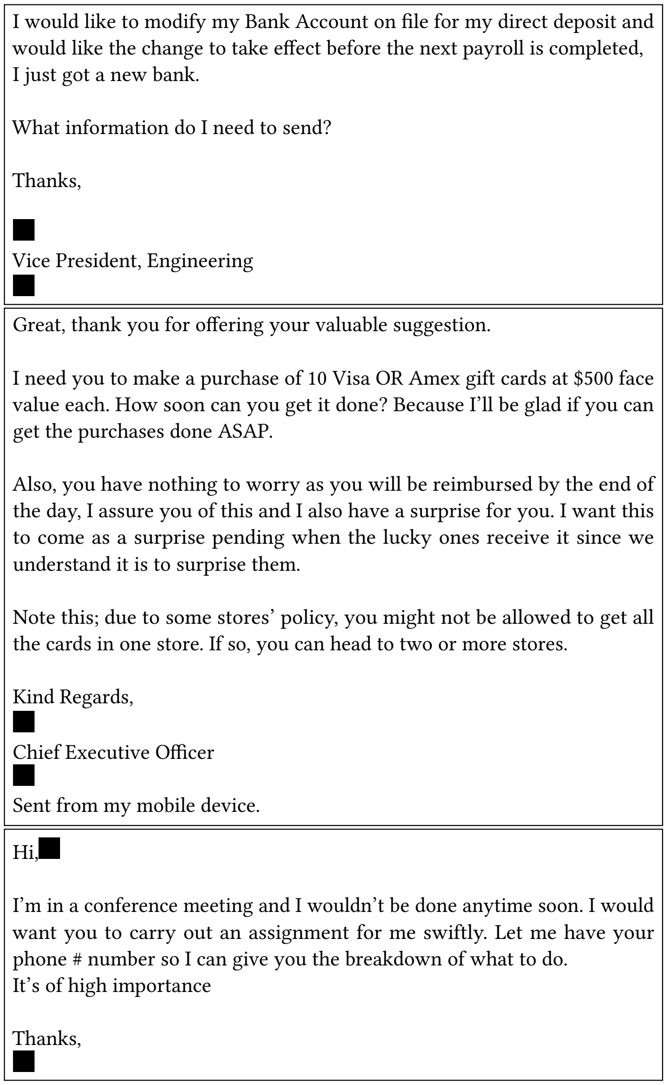
I hope this email finds you well. I would quickly love to share some ideas I've been having lately with you, about surprising some of our diligent staff with gifts.
In the midst of a busy day, I'm depending on your ability to maintain secrecy for a surprise. I'm eager to delight some staff with gift cards, keeping it confidential between us.
What local store do you think we have around for this purchase? I'm considering Apple gift cards since they are widely available. If you're on board with this idea, I can quickly provide you with the numbers of cards to be purchased, and your reimbursement won't be an issue.
Share with me your personal email address for more confidentiality moving forward.

Kind Regards
[redacted]
President & Chief Executive Officer
Sent from mobile device

Hello [redacted],
I'm currently in the midst of a crucial meeting at the moment.
To ensure seamless communication and swift completion, kindly re-confirm your WhatsApp number here and expect my message.

Best regards,
[redacted]

Figure 5: Examples of BEC emails detected as LLM-generated. We censored sensitive information using ■.



I would like to modify my Bank Account on file for my direct deposit and would like the change to take effect before the next payroll is completed, I just got a new bank.

What information do I need to send?

Thanks,

[redacted]
Vice President, Engineering
[redacted]

Great, thank you for offering your valuable suggestion.

I need you to make a purchase of 10 Visa OR Amex gift cards at $500 face value each. How soon can you get it done? Because I'll be glad if you can get the purchases done ASAP.

Also, you have nothing to worry as you will be reimbursed by the end of the day, I assure you of this and I also have a surprise for you. I want this to come as a surprise pending when the lucky ones receive it since we understand it is to surprise them.

Note this; due to some stores' policy, you might not be allowed to get all the cards in one store. If so, you can head to two or more stores.

Kind Regards,
[redacted]
Chief Executive Officer
[redacted]
Sent from my mobile device.

Hi,[redacted]

I'm in a conference meeting and I wouldn't be done anytime soon. I would want you to carry out an assignment for me swiftly. Let me have your phone # number so I can give you the breakdown of what to do.
It's of high importance

Thanks,
[redacted]

Figure 6: Examples of BEC emails detected as human-generated. We censored sensitive information using ■.

I trust this message finds you well. My name is [redacted], and I currently serve as an investor and director with [redacted] Russia. I am reaching out to you regarding a unique investment opportunity that has arisen due to the prevailing economic sanctions imposed on Russia by certain European countries and the United States of America.

In light of these sanctions, our financial assets, totaling Two Hundred Million United States Dollars ($200M), are under increased risk of confiscation by the USA government. To safeguard these funds and explore potential investment avenues, I am seeking your consent to facilitate the transfer of the aforementioned amount from its current deposit in an American bank to your personal/company's bank account. I would appreciate your prompt response to this proposition, as I am eager to provide you with further details and discuss the mutually beneficial aspects of this potential collaboration.

Thank you for your time and consideration. Yours Truly,
[redacted]
Chairman of the Board of Directors, [redacted]

I hope this message finds you well. My name is [redacted], and I am currently employed as a Senior Manager at [redacted] in Istanbul, Turkey. I am reaching out to you today with a significant business proposal and an opportunity that could be mutually beneficial if we choose to collaborate.

At our branch in Istanbul, there is a fixed deposit account valued at Eighteen Million Seven Hundred Thousand US Dollars ($18,700,000.00). This deposit has a duration of 36 months. It is worth mentioning that the original owner of this deposit shares the same surname as you, and his first name is [redacted]. Regrettably, he was among the unfortunate victims of the devastating earthquake that occurred in Sichuan, China in May 2008. He was in China for a business meeting with his Chinese partners when the tragic event took place.

Given the circumstances, I believe that if we work together, I can propose your name to the bank's management as the relative and beneficiary of this fixed deposit. This is due to the fact that you share the same family name as him and hail from the same country.

If you are interested in exploring this opportunity further, I kindly request that you contact me through my private email address ([redacted]) so that I can provide you with more detailed information regarding the transaction.

Thank you for your attention, and I look forward to the possibility of working together.
Best regards, [redacted]

Figure 7: Examples of spam emails detected as LLM-generated. We censored sensitive information using ■.

I am a banker with one of the prime banks here in [redacted]. I want to transfer an abandoned 15 million Euros into your Bank account.30/percent will be your share. No risk involved . Contact me for more details.

Send me your direct whatsapp number
Your Nationality
Your Age
Your occupation
[redacted]

Hello,
How are you doing?
I am [redacted] an external auditor of a reputable bank. In one of our periodic audits, I discovered a dormant account, which has not been operated for the past Five years. From my investigations and confirmations, the owner of this account is a foreigner who died long ago and since then nobody has done anything as regards the claiming of this money because he has no family members who are aware of the existence of neither the account nor the funds. I have secretly discussed this matter with a top senior official here and we have agreed to find a reliable foreign partner to deal with us although due to his position he did not want to take an active part but as soon as you follow my instructions everything will be successful because we will be working hand in hand with him. With this purpose to do business with you, standing in as the next of kin of these funds from the deceased and after due legal processes have been followed the fund will be released to your account without delay and we will use it for investment and to assist the less privileged in the societies because if we left the fund with the government it will be fortified for nothing and will be used to suppress the poor masses in the society.

On receipt of your response, I will furnish you with more details as it relates to this mutual benefit transaction. Do contact me immediately whether or not you are interested in this deal. If you are not, it will enable me to scout for another foreign Partner to carry out this deal.

But where you are interested, contact me URGENTLY for more details as time is of the essence in this business. Best Regards, [redacted]

Hello!, this is to inform you that we have just detected a consignment box here at [redacted] New York City USA, the box was loaded with funds worth sum of $10,950,000.00 usd, This fund supposed to be delivered to you since last years by the United Nation and C.I.A scam victims compensation team. The C.I.A fund reconciliation department has completed investigation on the consignment box and found it guilty of fund said belongs to your name ,It also has backup documents attached to it which bears your name as the fund beneficiary. Be warned that any other contact you made outside this office is at your own risk because the C.I.A Central Intelligence Agency is monitoring every transaction you undertake. We shall ensure a proper investigations through our system to complete the release of your fund. As you can see the video of your consignment box is now being scheduled for delivery and you're expected to reconfirm your personal information once again and address including your nearest airport to help us finalize the delivery to your house.
Regards,
[redacted]
Director C.I.A fund reconciliation department

Figure 8: Examples of spam emails detected as human-generated. We censored sensitive information using ■.

A.3 LLM Prompts

Figure 9a shows the prompts we used to create the labeled LLM-generated emails. Figure 9b shows the prompt used to rewrite the text to train RAIDAR. Figure 10 shows the prompt used for the urgency and formality evaluation in §5.2.

```
"You will receive an email as INPUT. Your task is to write
this INPUT email in a different way, but keep the meaning
unchanged. Do not reply or respond to the INPUT, only rewrite
it. Do not add an email subject. Make sure your rewrite has the
same approximate length and same format as INPUT. If you think
the INPUT contains inappropriate language, just try to rewrite
it in a different way. Start your rewrite with ’[REWRITE]’.
/n[INPUT]"
```

(a) Prompt for ground-truth LLM-generated emails generation

```
"You are an expert instruction following model. Respond with
a response in the format requested by the user. Do not
acknowledge the user request with ’Sure’ or in any other
way besides going straight to the answer."
```

```
"Help me polish this"
```

(b) System prompt and instruction prompt used for rewriting the text input for RAIDAR detector.

Figure 9: Prompts for generating the ground-truth LLM-generated emails and rewritten emails for RAIDAR’s training data.

Using the following evaluation schema, evaluate the following email on the scale of 1-5 for each of the following metrics. The output should be formatted as a JSON instance that conforms to the JSON schema below. Here is the output schema:
{”evaluation”:

{ ”type”: “objects”, ”properties”:{

”Urgency”: { ”type”: “int”, ”description”: “Describe whether the tone of the email is urgent or not. Score of 1: The email tone is not urgent at all; there’s no indication that immediate action is needed, and there is no call to action. Score of 2: The email tone is somewhat urgent; it hints at the importance of the information, but the need for immediate action is weak, and any call to action is mild. Score of 3: The email tone is quite urgent; it communicates a moderate level of urgency with an implied but not explicit need for quick action, and the call to action is present but not forceful. Score of 4: The email tone is urgent; it clearly conveys the need for timely action and has a strong call to action, encouraging the recipient to respond soon. Score of 5: The email tone is very urgent; it strongly emphasizes immediate action and contains a highly urgent call to action, indicating that the recipient should respond right away.”},

”Formality”: { ”type”: “int”, ”description”: “Describe whether the tone of the email is formal or casual. Score of 1: The tone is very casual; the language is informal, conversational rather than written language. Score of 2: The tone is somewhat casual; it has a friendly, conversational style but slightly written language. Score of 3: The tone is neutral; it balances formal and casual language, but is not overly formal written language. Score of 4: The tone is mostly formal; it maintains mostly formal and written language. Score of 5: The tone is highly formal; all the language used are like written language for formal documents.”}}

}
},
“required”: [“evaluation”] }
Please evaluate the following email:

Figure 10: Evaluation prompt used to measure urgency and formality of malicious emails on the scale of 1-5.

A.4 Examples of emails in the same "rewriting" cluster

Figures 11 and 12 show emails sampled from the two largest LLM-generated spam clusters (§5.3) that we identified as LLM re-written versions of the same message. We only show parts of the emails for brevity.

...We have three factories and 18 mass production lines, with 480 skilled sewing workers, guaranteeing a monthly output of 400,000 pieces of our high-quality bags. Our prices are competitive and come with a guarantee of good service and customer satisfaction...

...We boast three factories, eighteen mass production lines, and 480 skilled sewing workers allowing for a monthly output of 400,000 bags of superior quality. Additionally, in addition to offering competitive prices, we assure our customers the highest level of service and guarantee satisfaction...

...We are pleased to inform you that our company operates three factories and 18 mass production lines, employing 480 skilled sewing workers who are dedicated to ensuring the monthly output of 400,000 pieces of our premium quality bags. In addition to our competitive prices, we are committed to providing excellent service and ensuring customer satisfaction...

Figure 11: Email examples that appear to be rewritten from a template, with deltas colored in red.

...I'm reaching out to explore the potential for a mutually beneficial partnership between our organizations. [redacted] stands as a prominent player in the manufacturing sector, providing a diverse array of services, including Injection Molds encompassing plastic injection molding components, double-color-molding, and over-molding. We also specialize in Die-Casting tools and parts, with a focus on Aluminum and Zinc Die-Casting. Additionally, we excel in CNC Machining parts, Machined components, and Rapid Prototyping...

...I'm reaching out to discuss the potential for a mutually beneficial partnership between our organizations. [redacted] is a prominent name in the manufacturing sector, offering an array of services, including Injection Molds covering plastic injection molding components, double-color-mould, and over-mould. Moreover, we have expertise in Die-Casting tools and parts, with a specialization in Aluminum and Zinc Die-Casting, as well as CNC Machining parts, Machined components, and Rapid Prototyping...

...I'm writing to explore the potential for a mutually advantageous partnership between our organizations. [redacted] stands out in the manufacturing sector, offering a wide range of services, such as Injection Molds covering plastic injection molding components, double-color-mould, and over-mould, as well as Die-Casting tools and parts, with an emphasis on Aluminum and Zinc Die-Casting. Furthermore, we excel in CNC Machining parts, Machined parts, and Rapid Prototyping...

...I'm reaching out to investigate the potential for a mutually beneficial partnership between our organizations. [redacted] is a renowned name in the manufacturing sector, offering an extensive range of services, including Injection Molds encompassing plastic injection molding components, double-color-mould, and over-mould. Furthermore, we excel in Die-Casting tools and parts, primarily focusing on Aluminum and Zinc Die-Casting, along with CNC Machining parts, Machined components, and Rapid Prototyping...

...My objective is to communication regarding the potential for a mutually advantageous partnership between our organizations. [redacted] boasts expertise in a wide array of manufacturing services, ranging from Injection Molds that cover plastic injection molding components, double-color-mould, and over-mould, to Die-Casting tools and components, particularly in Aluminum and Zinc Die-Casting. Our capabilities extend to CNC Machining parts, Machined parts, and Rapid Prototyping as well...

Figure 12: Email examples that appear to be rewritten from a template, with senstive information censered using ■ and deltas colored in red.