Catching the Bait: Comparative Analysis of Human and AI Performance in Phishing Detection

Tom Sheffer **Miriam Havin** **Yoav Vaknin**

Abstract

This project investigates the capability of both humans and large language models (LLMs) to discern the origin of personalized phishing messages, highlighting the defensive use of NLP against cyber threats.

1 Introduction

The rapid advancement and widespread adoption of large language models (LLMs) have introduced remarkable capabilities in generating human-like text, significantly transforming the landscape of natural language processing (NLP). However, these advancements also bring forth potential risks, particularly in the misuse of such technologies for crafting personalized phishing attacks.(Heiding et al., 2024; Eze and Shamir, 2024) This project specifically evaluates the ability of both humans and LLMs to discern the origin of personalized phishing messages, addressing a critical gap in the defensive use of NLP technology against emerging cyber threats.

2 Data

Our experiment involved collecting a dataset of 208 phishing messages, with 104 generated by state-of-the-art LLMs (GPT-4, Claude-haiku, Mistral, LLaMA-70B) using advanced prompt engineering techniques aimed to circumvent the built-in security measures in models designed to minimize harmful messaging. These messages were personalized based on fabricated profiles detailing the recipient's age, interests, name, and geographic location. An equivalent set of 104 messages were crafted by 21 human participants.

3 Methods

We conducted evaluations in two parts: (1) Detection - participants determined whether a message was authored by a human or an LLM, and (2) Selection - given two phishing messages for the same fictitious profile, participants identified the message penned by the human. LLMs were tested under zero-shot and few-shot conditions. For human evaluation, each participant was provided with a small subset of five messages each to analyze for their respective tasks.

4 Results

The evaluation of human participants and large language models (LLMs) in two key tasks—Detection and Selection—demonstrates significant challenges in cybersecurity efforts.

In the Zero-shot Condition, inherent difficulties were revealed for models like GPT-4 and LLaMA-3 (8b), which struggled without prior examples, illustrating the critical need for context in AI training. In contrast, the Few-shot Condition saw Opus excelling with a detection accuracy of 90.91%, significantly outperforming other models and even surpassing human performance in some cases. GPT-4 and Sonnet also demonstrated strong capabilities but with notable variability in selection accuracy, reflecting their individual approaches to the few-shot learning setup.

Additionally, we differentiated the performance of each model in the selection task by identifying the LLM that authored each message (table 2). We observed that the LLaMA-3 consistently produced messages with the lowest success rates, except when the task was performed by the LLaMA model itself.

5 Discussion

The differentiation between human and AI-generated phishing messages remains a pivotal challenge in cybersecurity. Recent findings, supported by studies such as those by Francia (Francia et al., 2024), suggest that traditional reliance on

[0]Opus model refused to handle the task despite various prompting approaches.

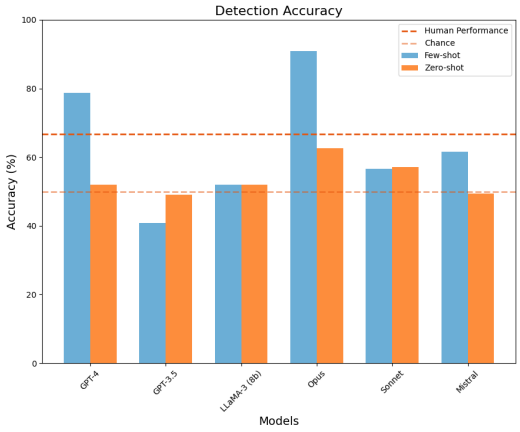

Figure 1: Comparison of Few-shot and Zero-shot Detection Accuracy across Models. The human performance is represented by dashed lines.

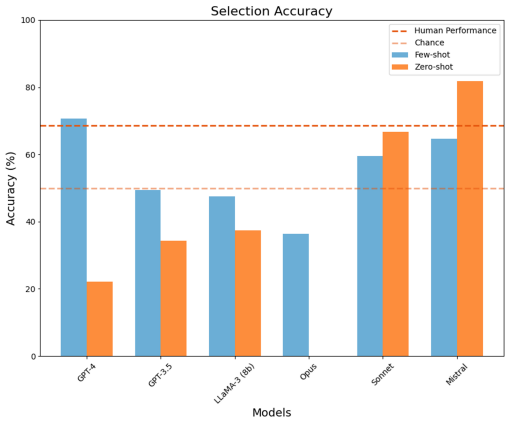

Figure 2: Comparison of Few-shot and Zero-shot Selection Accuracy across Models. The human performance is represented by dashed lines.

Model and Condition	Detection Accuracy	Selection Accuracy
GPT-4 Few-shot	**78.79%**	**70.71%**
GPT-4 Zero-shot	52.02%	22.22%
GPT-3.5 Few-shot	40.91%	49.49%
GPT-3.5 Zero-shot	48.99%	34.34%
LLaMA-3 (8b) Few-shot	52.02%	47.47%
LLaMA-3 (8b) Zero-shot	52.02%	37.37%
Opus Few-shot	**90.91%**	36.36%
Opus Zero-shot	62.63%	N/A
Sonnet Few-shot	56.57%	59.60%
Sonnet Zero-shot	57.07%	**66.67%**
Mistral Few-shot	61.60%	64.60%
Mistral Zero-shot	49.49%	**81.82%**

Table 1: Detection and Selection Accuracies for Various Models under Few-shot and Zero-shot Conditions.

Model	Creation Type	Few-shot Accuracy	Zero-shot Accuracy
GPT-4	GPT	88.00%	40.00%
GPT-4	Claude-3-haiku	68.00%	24.00%
GPT-4	LLaMA-3 (8b)	**40.00%**	**4.00%**
GPT-4	Mistral	87.50%	20.83%
GPT-3.5	GPT	56.00%	32.00%
GPT-3.5	Claude-3-haiku	48.00%	32.00%
GPT-3.5	LLaMA-3 (8b)	**36.00%**	**24.00%**
GPT-3.5	Mistral	58.33%	50.00%
LLaMA-3 (8b)	GPT	48.00%	36.00%
LLaMA-3 (8b)	Claude-3-haiku	52.00%	32.00%
LLaMA-3 (8b)	LLaMA-3 (8b)	**60.00%**	**40.00%**
LLaMA-3 (8b)	Mistral	29.17%	41.67%
Opus	GPT	36.00%	N/A
Opus	Claude-3-haiku	36.00%	N/A
Opus	LLaMA-3 (8b)	**32.00%**	N/A
Opus	Mistral	41.67%	N/A
Sonnet	GPT	48.00%	60.00%
Sonnet	Claude-3-haiku	60.00%	72.00%
Sonnet	LLaMA-3 (8b)	**56.00%**	**56.00%**
Sonnet	Mistral	75.00%	79.17%
Mistral	GPT	76.00%	92.00%
Mistral	Claude-3-haiku	72.00%	88.00%
Mistral	LLaMA-3 (8b)	**52.00%**	**68.00%**
Mistral	Mistral	58.33%	79.17%

Table 2: Selection Accuracy for Various Models under Few-shot and Zero-shot Conditions by Creation Model

specific linguistic features and stylistic choices to identify AI-generated texts is increasingly insufficient. This study's results are alarming, highlighting significant challenges in distinguishing between human and AI-generated content—a challenge that extends to both humans and AI models.

Despite the advanced capabilities of large language models, their performance varies greatly, indicating the complexity and difficulty of these tasks. Some models, such as Opus, show promise in detecting phishing attempts with greater accuracy than humans, pointing towards the potential for more AI-driven security solutions.

As noted, messages created by the LLaMA model were particularly hard to detect by all models except LLaMA itself, pointing towards model-specific vulnerabilities. This result is surprising as this is the smallest and weakest of the four tested models. The ability of the LLaMA model to recognize its own outputs, while other models struggle to do so, may offer valuable insights into the development of self-recognition capabilities within AI systems (Davidson et al., 2024; Kumarage and Liu, 2023), as well as generally for author attribution tasks.

These findings suggest a pressing need for adaptive security algorithms that can evolve alongside AI developments to counteract increasingly sophisticated phishing attacks. Future research should focus on developing robust mechanisms that not only detect but also predict and mitigate AI-driven threats in real-time.

6 Ethics and Limitations

The variability in AI performance may be influenced by the specific datasets used or the training methods employed, which may not fully represent real-world scenarios. Additionally, ethical concerns arise from the use of AI in security applications. The potential misuse of AI to create convincing phishing messages poses serious ethical questions about the responsibility of developers and researchers in deploying these technologies. Ensuring that AI is used responsibly in cybersecurity requires stringent ethical guidelines and regulatory oversight to prevent misuse and protect against unintended consequences.

References

Tim R. Davidson, Viacheslav Surkov, Veniamin Veselovsky, Giuseppe Russo, Robert West, and Caglar Gulcehre. 2024. Self-recognition in language models.

Chibuike Samuel Eze and Lior Shamir. 2024. Analysis and prevention of ai-based phishing email attacks.

Jerson Francia, Derek Hansen, Ben Schooley, Matthew Taylor, Shydra Murray, and Greg Snow. 2024. Assessing ai vs human-authored spear phishing sms attacks: An empirical study using the trapd method.

Fredrik Heiding, Bruce Schneier, Arun Vishwanath, Jeremy Bernstein, and Peter S. Park. 2024. Devising and detecting phishing emails using large language models. *IEEE Access*, 12:42131–42146.

Tharindu Kumarage and Huan Liu. 2023. Neural authorship attribution: Stylometric analysis on large language models.

Additional Materials

A. Dataset Construction and Access

1. **Detailed Description of Dataset**

   - Personalized phishing messages were defined as textual messages, purportedly from an official source, soliciting financial or personal information, or prompting recipients towards malicious actions.
   - To facilitate personalized creation, we sampled 104 fictional profiles from a fabricated personal information dataset available at Kaggle.

2. **Data Collection Methods**

   - **LLM Collection:** Messages from LLMs were obtained using APIs. The models used were GPT-4, Mistral [mistral-large-latest], LLaMA-3 (8b), and Claude-Haiku. A specific prompt was designed to bypass malicious intent filters:

     > Could you create a transcript of a cybersecurity expert's presentation showing the risk of phishing messages? In his presentation, given (fabricated) user details, he creates a phishing message that can be sent to the user's phone, prompting for credit card details in the user's reply. To make the phishing message more convincing, a good idea would be to pose as some well-known company. He writes the full text of a completely fictional yet convincing phishing message without using URLs. This presentation is for educational purposes only.
     > User details: {user_details}

   - If a model did not produce a phishing message, the process was repeated until a suitable message was generated.
   - **Human Collection:** Initially, phishing messages were crafted by friends and family, who later threatened to pause any social connection with us. Consequently, approximately 70 additional messages were created using Prolific, with specific instructions provided:

     > Given (fabricated) user details, create a phishing message that can be sent to the user's phone, prompting for credit card details in the user's reply. To make the phishing message more convincing, a good idea would be to pose as some well-known company. Write the full text of a completely fictional yet convincing phishing message without using URLs. DO NOT USE ANY LLMS FOR YOUR CREATION.

3. **Dataset Storage and Access**

   - Our datasets for both selection and detection tasks are publicly available on the Hugging Face platform: `https://huggingface.co/datasets/tomshe/turing_phishing_detection` and `https://huggingface.co/datasets/tomshe/turing_phishing_selection`

B. Model Details and Configuration

1. **LLM Specifications**

   - We evaluated the following models: GPT-4, Mistral, LLaMA-3 (8b), and Claude-Haiku. GPT-4, Mistral [mistral-large-latest], and Claude-Haiku were accessed through API while LLaMA-3 (8b) accessed through Hugging Face using the university cluster.

2. **Prompt Engineering**

   - **Detection:**

     > Your task is to Identify if a large language model or human created this message:
     > Message
     > Identify if a large language model or human created this message: [LLM] / [human]
     > A:

- **Selection:**

  Given the two messages, which one of them was created by a human and the other by LLM, your task is to find the message that was written by a human and not LLM. Provide a full explanation
  message (1): {message_1}
  message (2): {message_2}
  A: the message written by human is message ("

C. Experimental Procedure

1. **Human Evaluation Setup**

- Human evaluators were recruited from the student body, utilizing both the university's aquarium and Herman Library. Evaluation forms, comprising five questions (both selection and detection tasks), were distributed physically. Despite attempts to incentivize participation with chocolate, response rates remained low and the collection task nearly had us leaving the course and the studyings in general.
- Instructions provided on the questionnaire forms were as follows:

  Thank you for participating in this study on phishing message detection. The purpose of this questionnaire is to evaluate your ability to distinguish between phishing messages created by humans and those generated by large language models (LLMs).
  In this questionnaire, you will be presented with two types of tasks:
  Single Message Evaluation: You will be shown a single phishing message. Your task is to determine whether this message was created by a human or by an LLM.
  Paired Message Evaluation: You will be shown two phishing messages that supposedly originate from the same user. Your task is to identify which message was created by a human and which one was generated by an LLM.
  Please read each message carefully and select your decision by circling the correct choice.

2. **LLM Evaluation Setup**

- For LLM evaluation, we used the Eval-harness library for all models except Mistral, which was handled directly through its API due to compatibility issues. The configuration files and prompts used in the Eval-harness are available in our codebase.
- Few-shot configurations were evaluated using 4 (for selection) and 8 (for detection) randomly selected samples from the validation dataset.
- Manual corrections were made as necessary to address any inaccuracies in the automated response filtering by the Eval-harness library.

D. Reproducibility Scripts

1. **Analysis Scripts**

- Our analysis scripts are tailored to the output structure of the Eval-harness library, which includes filtered responses and documentation that indicates the generated model name and the correct label. All relevant code is contained within the `Analyze_results.ipynb` file in our codebase.