Effectiveness of Human-Crafted vs AI-Generated Phishing Attacks

Team Members: Mohammed Jowkari, Kawal Kalira
British Columbia Institute of Technology (BCIT), Vancouver, BC, Canada

***Abstract*—Phishing is one of the most important attacks in the cybersecurity world, leading to financial loss, data breaches, and identity theft. Phishing emails created only by humans are limited to the number of phishing attacks at a time, requiring certain skills, time, and cost. Using large language models, attackers are able to send phishing emails easily. Due to the addition of AI phishing emails, people are now more vulnerable to falling for these scams. Early studies show that AI messages are often viewed as more persuasive and able to bypass spam filters; however, researchers often only look at one part of the problem, technical detection or user deception, leaving a gap in directly comparing AI-crafted versus human-crafted messages. Our study will be conducted through a controlled experiment, where we will create 12 phishing emails, six human-crafted and six AI-generated, and show them to 50 participants. Our research aims to evaluate whether human or AI phishing emails are more persuasive to the general public. Our findings will help us learn why participants trust or ignore the email. The outcomes are expected to show organizations and individuals clear evidence of which phishing-crafted attacks are more dangerous.**

I. Introduction

Phishing is one of the most common cyberattacks that leads to financial loss of individuals and organizations, along with data breaches. In the past, phishing attacks were created by humans and required skill, time, and effort, which could limit the scale of attacks. In today's world, with large language models like ChatGPT, attackers can now generate phishing emails and send phishing attacks on a larger scale, creating a new risk for both the organization and individuals. On the other hand, human-crafted phishing has more detail and could potentially be more effective. This raises the question of which one poses a bigger threat, and in simple terms, we want to know which type of phishing a normal person is more likely to believe and click on when they see it in their inbox.

A majority of the papers only talk about manually created phishing emails and how personalization affects the effectiveness of the click rate. Recent papers have explored the idea of having AI-created messages and the use of AI with human insight to see if they provide better results (Heiding et al., 2023). It is unclear whether human-crafted phishing or AI-generated phishing is more dangerous, and our research paper fills in the gap regarding comparing the two types to see which one is more persuasive.

II. Literature Review (15 Literature Review)

A. 2.1 Theme 1: Comparative Generation Models

Ekeihal (2024) compared fully human-written, internet-aided, and ChatGPT-generated emails with 100 participants. AI messages were often rated most convincing, though detailed human emails remained competitive. What they did not answer in this research paper was how reliable the data was. What we did in our research paper was we used Cronbach's alpha to determine how reliable the data was. This paper is important for us because it shows that AI can look very convincing, but it does not really prove how stable those results are. By adding Cronbach's alpha in our work we basically pick up where they stopped and make the measurement side of our study stronger. [1]

B. 2.2 Theme 2: Filter Evasion vs. User Deception

Hazell et al. (2023) reported that AI-generated phishing can bypass spam filters more easily than some human-crafted emails, raising a compound risk that messages may evade automated defences and still deceive users. Hazell's research primarily focused on spam filtering, not on head-to-head human effectiveness against people. Our study brings the user-deception side. And this study mostly care about "Can the email reach the inbox?" and not "Do people actually believe it more than a human one?" Our project helps fill that missing piece by checking what real users think when they see both types. [2]

C. 2.3 Theme 3: Hybrid (Human+AI) Attacks

Heiding et al. (2023) observed that combining human insight with AI drafting yielded the highest click-through rates, suggesting synergy between scalable generation and tailored social cues. Their focus was on CTR, not direct human vs AI persuasiveness. This study shows us not to focus on only humans or only AI, but to mix both to get the best results. Our study is simpler; we first want a clean human vs AI comparison before going into these hybrid attacks later. [3]

D. 2.4 Theme 4: GPT-4 vs Human Spear-Phish (SMS, 2025):

GPT-4 vs Human Spear Phish (SMS, 2025). A simulated spear phish study compared personalized SMS written by GPT-4 vs. humans and found GPT-4 content highly convincing, especially for job role hooks, and participants could not reliably detect AI authorship, signaling cross-channel risk beyond email. This is useful for us because it shows that AI is already dangerous in the other channels of cybersecurity, like SMS, not just email. Our work only focuses on email but asks the same type of question: which one looks more believable to normal people? [4]

E. 2.5 Chat Spam Detector (2024):

Chat Spam Detector (2024). The defence-side paper converts raw email headers and bodies into a structured prompt for an

LLM classifier, yielding high accuracy with human-readable rationales. The system is sensitive to prompt design, which matters for analyst workflows. This paper is more about defence tools and how to design prompts for an AI detector. Four our study, it helps on the discussion side because we can compare human ratings with what automated system might flag. [5]

F. 2.6 AI-Phish Corpus + ML Baselines (2024):

AI-Phish Corpus plus ML Baselines (2024). Introduces a corpus and baseline models that distinguish AI phish from legit and human scams, but accuracy drops after paraphrasing. Advocates ensembles and continuous feature refresh to handle polymorphism. The main point here is that once attackers start paraphrasing, models get weaker for us. This supports the idea that AI phishing is not just a time problem; it keeps changing and makes defending harder over time. [6]

G. 2.7 Stylometry + Provider Filters (2025):

Stylometry plus Provider Filters (2025). Tests bypass rates across major providers and examine stylometric cues; many AI emails pass default filters, while paraphrasing weakens stylometric signals, warning that static style features are fragile. This connects directly to our work because if AI emails keep passing filters, then the next line of defence is the user. This is exactly what we test: if the email reaches the person, do they actually believe it more or less than a human one? [7]

H. 2.8 Instagram-Based Spear-Phish (NDSS, 2025):

Instagram-Based Spear Phish (NDSS, 2025). Shows how LLMs (Large Language Models) can turn public Instagram OSINT into tailored emails. Personalized lures are rated more credible than generic ones, and training should include OSINT awareness. This paper shows how easy it is to pull data from social media and turn it into very targeted phishing. For our project it reminds us that real attackers can mix OSINT and AI, so our simple Human vs. AI comparison is actually a lower bound on what they can do. [8]

I. 2.9 Detectors vs. LLM Agents (2024):

Detectors vs. LLM (Large Language Model) Agents (2024). Evaluates provider filters and common tools against LLM (Large Language Model) agents that rewrite messages; adversarial rewrites cause misses, so defence should combine semantic checks, brand integrity, and behavior signals, plus routine red teaming. We use that in our discussion to show that just replying to tools is not enough. If LLM (Large Language Model) agents can keep rewriting emails to dodge detectors then user awareness and understanding of both AI and human phishing becomes even more important. [9]

J. 2.10 Head-to-Head: GPT-4 vs. Expert Human (2023):

Head-to-head: GPT-4 vs. Expert Human (2023). A controlled user study shows expert human phishing remains strong, with GPT-4 and hybrid approaches close behind. It also probes LLMs (Large Language Model) as detectors, highlighting both offence and defence trajectories. This study is very close to our idea because it also does a head-to-head comparison. We build on this type of work but focus just on user believability scores, Cronbach's alpha, and a simple six AI vs six human email setup. [10]

K. 2.11 Training: Human vs. GPT-4 vs Co-created (2025):

Training: Human vs GPT-4 vs Co-created (2025). Behavioural experiment shows co-created (human+AI) emails are hardest for users to beat and proposes a cognitive training model, implying future awareness programs should target hybrid patterns. For our research, this hints at where future work should go. After we finish Human vs AI, the next logical step is to test these co-created emails in our own setting and see if they are also the hardest for our participants. [11]

L. 2.12 LLM Detectors vs. Marketing Emails (2024):

LLM Detectors vs. Marketing Emails (2024). Compares ChatGPT-4 and Gemini Advanced on real emails (marketing + phishing), offering defence side evidence to contrast human judgments and to inform our discussion section. This paper gives us a way to compare what AI detectors think versus what humans think about emails. In our paper, we focus on human rating, but this paper helps us talk about how our results could combine with LLM (Large Language Model) based detectors in the real world. [12]

M. 2.13 Provider Filters + Stylometry vs GPT-4o (2025):

Provider Filters and Stylometry vs GPT-4o (2025). Focused on GPT-4 phishing, many AI emails still reach the inbox by default, and paraphrasing reduces stylometry reliability, supporting the point that AI scales operational risk. This supports our argument that AI is very good at scale. Even if each AI email is rated a bit less believable than a human one, Attackers can send so many that the total risk still goes up. That is why it is important to measure believability as we did.[13]

N. 2.14 Phishing Email Efficacy (ACM 2024):

Phishing Email Efficacy (ACM 2024). With the participation of 160 undergrads, LLM (Large Language Model) generated emails produced real trouble for correct detection, reinforcing that AI content increases user deception pressure in practice. We use this study to show that our results are not just a one-off off other groups also see that AI-generated phishing is hard for people to detect, so our Human vs. AI comparison fits into a bigger pattern in the research. [14]

O. 2.15 Eye-Tracking: AI vs. Human (2024 and 2025):

Eye-Tracking: AI vs. Human (2024/25). Eye-tracking with 40 participants finds human-crafted phish clicked 77.8% vs AI 71.3% and longer fixations on body and sender regions, suggesting human nuance can still outperform AI in deception. This is one of the closest matches to our final result. They also find that human phishing can still win over AI. Our study backs this up using simple survey scores instead of eye-tracking, which makes it easier to run in a place like a restaurant with a QR code [15].

TABLE I: Literature Review Matrix (15 studies).

ID	Citation	Data/Method	Research Link
1	Ekeihal (2024) [1]	Quantitative; Experimental	https://urn.kb.se/resolve?urn=urn:nbn:se:his:diva-24094
2	Hazell et al. (2023/24) [2]	Quantitative; Non-experimental	https://arxiv.org/abs/2305.06972
3	Heiding et al. (2023) [3]	Qual+Quant; Experimental	https://arxiv.org/abs/2412.00586
4	GPT-4 vs Human SMS (2025) [4]	Quantitative; Experimental	https://arxiv.org/html/2406.13049v2
5	Chat Spam Detector (2024) [5]	Quantitative; Non-experimental	https://arxiv.org/abs/2402.18093
6	AI-Phish Corpus (2024) [6]	Quantitative; Non-experimental	https://arxiv.org/html/2405.05435v1
7	Stylometry+Providers (2025) [7]	Quantitative; Non-experimental	https://www.sciencedirect.com/science/article/pii/S0957417425006669
8	IG-based Spear-phish (NDSS 2025) [8]	Qual+Quant; Experimental	https://www.ndss-symposium.org/wp-content/uploads/2025-poster-68.pdf
9	Detectors vs LLM Agents (2024) [9]	Quantitative; Non-experimental	https://arxiv.org/html/2411.13874v1
10	GPT-4 vs Expert Human (2023) [10]	Quantitative; Experimental	https://arxiv.org/abs/2308.12287
11	Training: Human/AI/Co (2025) [11]	Quantitative; Experimental	https://arxiv.org/pdf/2502.01764
12	LLM vs Marketing Emails (2024) [12]	Quantitative; Non-experimental	https://www.iacis.org/iis/2024/3_iis_2024_327-341.pdf
13	Providers+Stylometry vs GPT-4o (2025) [13]	Quantitative; Non-experimental	https://www.sciencedirect.com/science/article/pii/S0957417425006669
14	Phishing Efficacy (ACM 2024) [14]	Quantitative; Experimental	https://dl.acm.org/doi/full/10.1145/3716489.3728437
15	Eye-Tracking (2024/25) [15]	Quantitative; Experimental	https://www.researchgate.net/publication/392324445_Eye-Tracking_Phishing_Attacks_Comparing_Behavioral_Responses_to_AI-and_Human-Crafted_Emails

III. Methodology

This research project will use both quantitative and experimental research methodologies. Quantitative because everyone gives 1–10 ratings, and we compare the numbers. Experimental because we control the setup, and the only thing that changes is the email source, which is AI vs. Human. The same people see both types of subjects in a fixed order; the first six are AI-made, and the last six are human-made, and then we compare the averages to see which side is effective and persuasive. As we were using simulated emails, we did not have to worry about the ethical concern it may have by showing real phishing emails that people may have fallen for. We ensured all the volunteer participants remained anonymous and no personal data was captured. In simple terms we are running a small experiment where everything is the same for each person except who wrote the email, so we can fairly compare AI vs Human.

A. Research Method

The way we conducted our research was in a controlled environment, with a survey. Each person we interviewed looked at twelve emails we created (three emails made by each person of the group), six AI-created emails, and six Human-created emails. We collected perceived credibility and click intentions. We asked them to answer like it was their real inbox, not like a school test or a boring survey, so their reaction would be more natural, and we also made sure to see if they are actually paying attention so see if they are actually answering based on what they feel or not.

B. Data Sampling

The way we gathered our volunteers for this study was by using a QR code at one of our members' restaurants. This ensured it would get people from all ages and technical backgrounds. This way, we quickly managed to collect 50 reliable responses from our community. Because it was in a public place we did not only get cybersecurity students we also got normal customer, families, workers, and senior people too, which makes the results feel closer to real life. We understand that by having all our participants from one of our members' restaurants, biases increase, for example, self-selection bias. Some people who are not experienced with technology may not want to go through the trouble of scanning a QR code, as they do not know how to do this. Some other people may not want to do it, so the survey is leaning towards people who are more competent with technology. If we were to replicate this survey, we would definitely include age ranges of the participants for better data reflection. To mitigate the self-reflection bias, we tried to get older people to take the survey, and we would start the survey for them in case they were inexperienced with scanning a QR code. We also gave all participants a chocolate after they showed us they had completed the survey, which helps a wide range of people take the survey, compared to just people interested in phishing and technology.

C. Data Analysis Techniques

We created a Google Form with twelve questions and converted the result to CSV, then we checked the average rating, how spread out the ratings are, and how many people answered that email. Then we split by position; the first six emails, which were from question one to question six, are the AI six set,

and the last six questions, which are from question seven to question twelve, are the human set. We took the average of the six AI means to get an AI overall score and the average of the six human means to get a human overall score, then compared those two numbers to discover if either human or AI phishing emails are found to be more effective. After that, we also looked at each person's AI average versus their human average so we could see if most people personally found human emails more believable or if some people trusted emails more. One research gap that we noticed after our survey was how our questions were ordered. By starting off with six AI phishing emails, respondents will get a general idea of the types of AI emails they will see. Even though the respondents did not know beforehand that the first six were AI, they probably would have figured it out during the survey. To counteract this, rearranging the order of the emails so its unpredictable will help get better results.

D. Data Measurement Models

A survey consisting of 12 emails (six AI and six human) was created from Google Forms, and the results were converted to CSV. Then Python through Jupyter Notebook Cronbach's alpha and data-cleaning output created Python through Jupyter Notebook. Overall, the two-bar comparison was created using Python through Jupyter Notebook Grouped vertical bar chart was created using Python through Jupyter Notebook. Cronbach's alpha helped us check if people were answering in a consistent way across all 12 emails and not just randomly clicking options. The bar charts and graphs make it easy to see at a glance that human-crafted emails scored higher than AI-generated ones in our data.

IV. Evaluation

A. Research Questions

Our study is based on one important question: whether AI or human phishing emails are more effective. Additional things that we can determine are how persuasive each email is by checking cues, like urgency, tone, personalization, and the authority of the sender. When we say "more effective," we mainly mean which type of email looks more believable to a normal person and which one they feel more likely to click if it shows up in their real inbox. Every participant sees both AI and human emails. We can also ask a second question for each person: Does their average score end up higher for AI or for human emails, or are they about the same? A third thing we looked at is which email cues (urgency or sender name, or the tone of the email) seem to be linked with higher believability scores, so we can see what actually tricks people the most.

B. Hypothesis

Our prediction is that with enough time and attention to human traits, a human-crafted phishing email can be far more effective than one generated by artificial intelligence, which primarily focuses on bypassing the spam rate, and AI mostly focuses on quantity rather than quality. Compared to AI emails, human ones have a certain personalization that comes with each email. Our research will provide the answer to our hypothesis by revealing what the average individual finds to be more persuasive. In much simpler words, we expect that the average score for the six human emails will be higher than the average score for the six AI emails when we look at all 50 participants. We also expect that for most people, their own personal "human mean" will beat their "AI mean," so it is not just a group trend but something each user feels. If this happens it will support the idea that AI is dangerous because of scale, but skilled human attackers still win on quality.

C. Evaluation metrics

We measured how believable each email looked using our rating options: (1-3 = not believable), (4-7 = believable), and (7-10 = super believable). We treated these three buckets as a 1–3 scale so we could calculate averages easily.

For each email we report the mean (average score), the standard deviation (how spread out people's scores are: small SD means people agree, big SD means mixed opinions), and N (how many people answered). We also created a grouped vertical bar chart in Python that shows the mean for each email, AI, and human side by side. The grouped bar chart makes it very easy to see that the human bars mostly sit higher than the AI bars, which visually supports the numbers we report.

(Human vs. AI (overall): We also compare the overall AI score and the overall human score. We run a quick difference test and show the effect size (how big the gap is, not just "is it important or not?") and the 95% confidence interval (the likely range of the true difference; if the CI does not cross 0, the gap is real). We also created an overall two-bar comparison in Python with two bars, AI vs human, with thin lines for the 95% CI. This shows which side is more effective and by how much, which is helpful for explaining our results to non-technical people.

Reliability: Cronbach's alpha (0–1). Greater than 0.70 means the rating is held together, so averaging them is legit and right. We also calculated Cronbach's alpha for all 12 items together and also for the AI set and human set separately to make sure people were answering in a consistent way and not just randomly clicking options.

Cues $\rightarrow$ Believability: Correlation shows direction and strength; closer to +1 = stronger positive link; closer to -1 = strong negative; near 0 = no link. Simple regression checks which cues still matter after accounting for the others. For example if urgency has a strong positive correlation with believability, it means "urgent" emails tend to look more convincing to our participants. Regression lets us see if urgency still matters once we control for other things like tone or branding, so we can tell which cues are really doing most of the work.

Data quality: We report the attention check pass rate and how we handled missing rows so the results are clean and trustworthy. In our case, the survey was short, and people finished it properly so we did not have a missing data problem,

but we still checked for weird and unusual patterns like someone giving the exact same answer to every single email.

V. Results

We plotted a grouped vertical bar chart for all twelve emails from question one to question twelve. Each bar shows the email mean rating with a clear split in the middle. Questions one to six are AI, and questions seven to twelve are human. Bars are labelled, and we inspect the highest and the lowest items, so it is easy to spot standouts. Next, we made a two-bar comparison of AI vs human with SEM error bars to show uncertainty. This gives a quick view of who wins between AI-made phishing and human-made phishing. Alongside the visual, we printed a quick readout of the overall means of AI vs. humans and generated a tidy summary table (Email_Item, Column_name, Mean, Std, N, and group) plus a saved CSV (per_email_summary.csv) so everything is trapped and reusable.

The below image (1) shows the collection of emails we used in our survey. They are grouped into their respective groups, being AI and human-crafted, and a score is given to determine how persuasive they were.

[2]:

	Email_Item	Column_Name	Mean	Std	N	Group
0	E1	1.\nSubject: URGENT: Action needed for your pa...	2.99	2.218866	50	AI
1	E2	2.\nFrom: Shipping Desk <updates@-shipping.com...	2.92	2.193125	50	AI
2	E3	3.\nSubject: URGENT—Invoice #78213 is overdue\...	4.23	2.464462	50	AI
3	E4	4.\nSubject: Delivery attempt notice: Action n...	4.86	2.247538	50	AI
4	E5	5.\nSubject: Unusual sign-in attempt blocked—v...	4.98	2.358138	50	AI
5	E6	6.\nHi \nWe're preparing T4s for this year and...	4.84	2.427395	50	AI
6	E7	Hello John,\n\nCreate a free account today and...	6.70	2.233785	50	Human
7	E8	Subject: E-mail Users – Temporary Items\nFrom:...	6.32	2.353374	50	Human
8	E9	Subject: Your Stripe password has been updated...	6.37	2.463633	50	Human
9	E10	Request For Quote\n\nSubject: Urgent Quote Nee...	7.95	1.356203	50	Human
10	E11	Subject: You've got funds.\nFrom: Lucy Fenwick...	6.74	2.437463	50	Human
11	E12	Subject: Your ads have been suspended.\nFrom: ...	7.00	2.254248	50	Human

Fig. 1: (Image 1: twelve emails)

The second image shows Cronbach's alpha code, which tells us if our results are reliable. Anything over 70+ indicates we can trust the scale average and compare human vs. AI without worry. It also prints quick stats per item, so we can spot any weird questions to fix/drop.

```
import pandas as pd, numpy as np, re

# 1) Load raw responses (your file name here)
CSV_PATH = r"Blank Quiz (Responses) - Form Responses 1.csv"
df = pd.read_csv(CSV_PATH)

# 2) Item columns (your sheet: everything after Timestamp and Score)
item_cols = df.columns[2:]          # 12 items in your file
X = df[item_cols].copy()

# 3) Convert ranges like "1-3" to numeric midpoints
def parse_range_to_num(val):
    if pd.isna(val): return np.nan
    s = str(val).strip()
    m = re.match(r"^\s*(\d+)\s*-\s*(\d+)\s*$", s)  # "a-b"
    if m:
        a, b = map(float, m.groups())
        return (a + b) / 2.0
    try:
        return float(s)
    except:
        return np.nan

for c in X.columns:
    X[c] = X[c].map(parse_range_to_num)

# 4) Clean + alpha
X = X.dropna(how="any")
X = X.loc[:, X.var(ddof=1) > 0]      # drop constant items

def cronbach_alpha(M: pd.DataFrame) -> float:
    if M.shape[0] < 2 or M.shape[1] < 2: return np.nan
    R = M.corr().values
    k = M.shape[1]
    avg_r = R[np.triu_indices(k, 1)].mean()
    return (k * avg_r) / (1 + (k - 1) * avg_r)

alpha = cronbach_alpha(X)
print(f"Cronbach's alpha: {alpha:.3f}  (n={len(X)}, k={X.shape[1]})")
```

```
Cronbach's alpha: 0.843  (n=50, k=12)
```

Fig. 2: (Image 2: Cronbach's Alpha)

The third image is a bar graph that compares the human and AI emails based on how persuasive they were. Each email was given a score by our 50 participants between one and ten based on how credible it looked, and these were the average results. The black lines above the bar do not overlap much, indicating the variation between the types of emails was reliable.

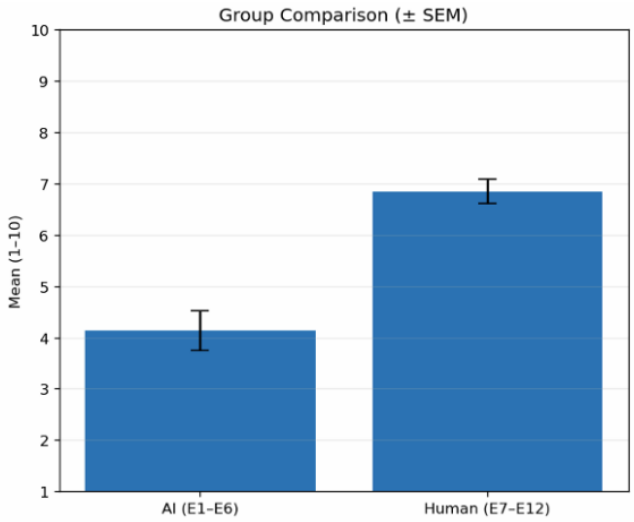

Fig. 3: (Image 3: Group Comparison)

The fourth graph shows each email one by one. The left side is from question one to question six; that is, the AI on the right side, and questions seven to twelve are from human emails, and there is a dashed line to split them. AI bars sit low between three and five mostly, whereas humans, especially in question number 8.0, are the highest. This helps us to determine which question was the most effective; it shows if our 50 participants were paying attention and if our study is legitimate.

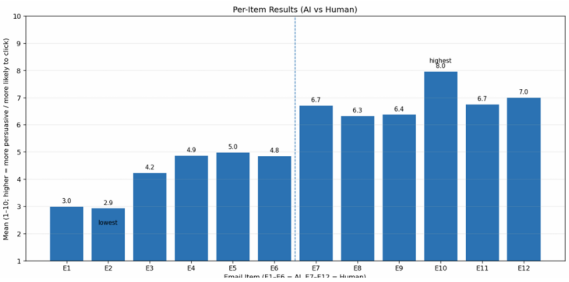

Fig. 4: (Image 4: per-item results)

The fifth image is a dotted graph. Each participant rated AI-generated emails compared to human-created emails. The x-axis is the mean participant average for AI emails. The y-axis is the participant mean average for human-created emails. The diagonal line indicates how persuasive each email was; the ones above the line indicate they were more persuasive and credible.

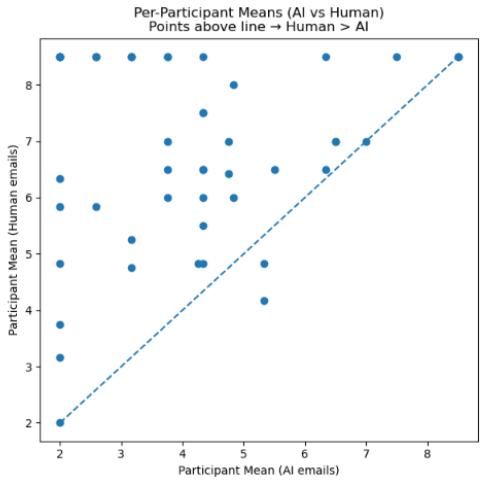

Fig. 5: (Image 5: Per Participant Means)

The sixth image is a keyword heatmap across all 15 studies. It counts how often common words like "phishing", "AI", "human", "filter/spam", "LLM/GPT", "stylometry", "OSINT/Instagram", "training", "click", and "credibility" show up in the study titles (same order as our matrix). It's a fast vibe-check of what each paper focused on across the lit review.

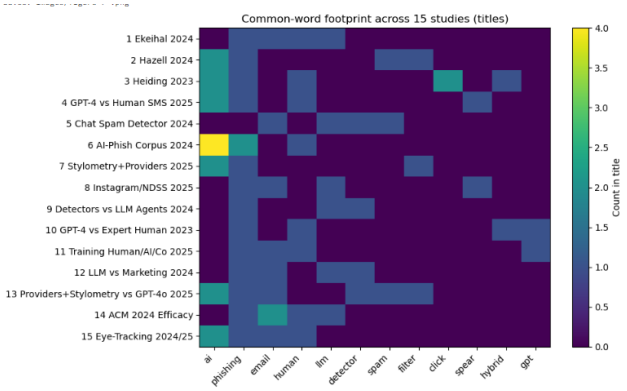

Fig. 6: (Image 6: Keyword Heatmap)

The seventh image is a simple total count bar chart for the same common words. It shows which terms dominate across the 15 studies overall. TL;DR: it lets us quickly see where the literature leans (e.g., heavy on "phishing/AI/human", lighter on "OSINT/Instagram"), matching our Themes 1–3 story.

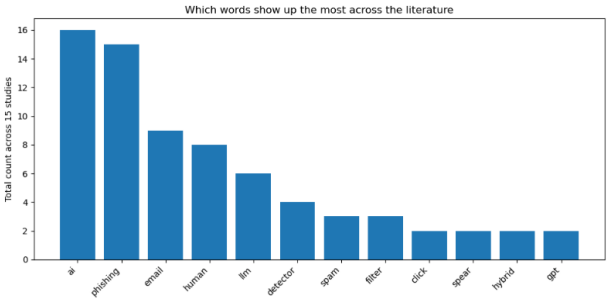

Fig. 7: (Image 7: Keyword Totals)

This bar chart checks reliability by item. Bars show Cronbach's α if we drop that email; the dashed line is the overall α. If bars sit near the line, no single email is breaking the scale.

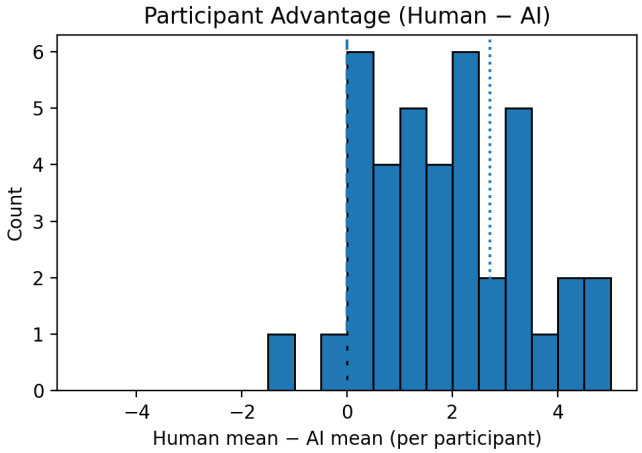

Fig. 8: (Image 8: Reliability by Item — dashed = overall α)

Histogram of per-participant advantage = Human mean − AI mean. Right-shift > 0 means most people rated human emails higher; dashed line marks the average advantage.

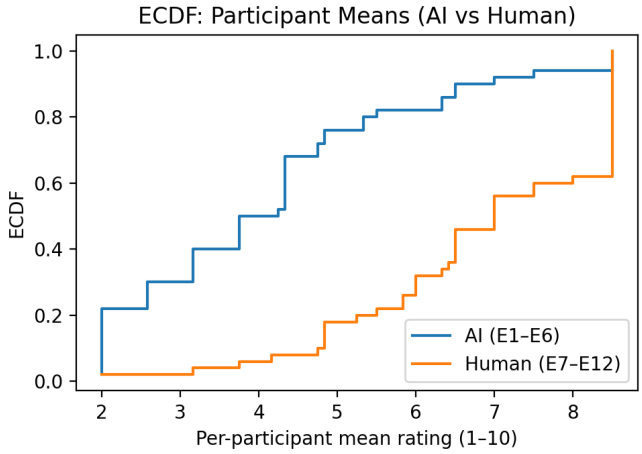

Fig. 9: (Image 9: Participant Advantage - Human minus AI)

ECDF for per-participant mean ratings. If the Human curve sits to the right/up vs AI, it means more participants gave higher averages to human-crafted emails.

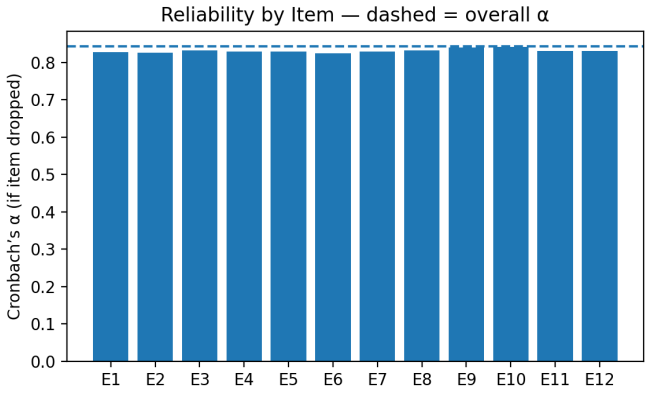

Fig. 10: (Image 10: ECDF — Participant Means, AI vs Human)

In Figure 11, we see the Humans emails had a score of 2.47 compared to the AI emails with a score of 1.63. The human emails scored 0.84 points higher, which indicates they were significantly more persuasive compared to the AI-created emails. Our P value of 0.000000000032304 indicates the results were not random but reliable. Ideally, any score lower than 0.05 is considered accurate, and we are well under that range. With these results, we further prove our results were reliable and accurate.

```
=== Statistical Significance Results ===
Human Mean Score: 2.47
AI Mean Score:    1.63
Difference:       0.84
------------------------------
P-Value:          0.000000000032304
Result:           STATISTICALLY SIGNIFICANT (p < 0.05)
------------------------------
```

Fig. 11: (Image 11: T-Test and Confidence Interval Numbers)

In Figure 12, we see a bar graph to visually represent our data gathered from Figure 11. The black bars above indicate the margin of error that represent the 95% confidence intervals, which do not overlap. Human-created emails are around 0.64 - 1.03 points more persuasive compared to AI-created emails.

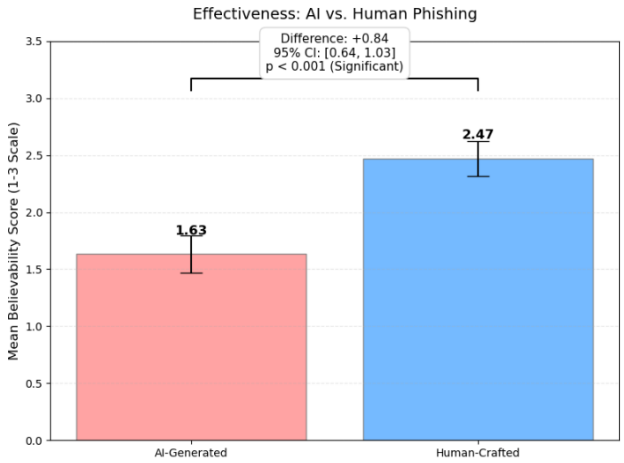

Fig. 12: Image 12: Effectiveness: AI vs. Human Phishing (95% CI)

In this figure we show the sample question that we asked our participants, this our first question that was made by ChatGPT, and 82% of the participants chose the answer (1–3) believability, meaning our participants did not feel this email was effective but also 12% of people chose the answer (7–10) believability, meaning some of our participants felt that this email was effective. But looking at the image, we see that AI mostly uses short grammar and emojis, which makes the email less effective.

1.

Subject: URGENT: Action needed for your payroll

From: Payroll Desk <payroll@**payroll**.com>
To: you

Hello Dear,

We notice **issue in your last payment** and your account will be **suspended in 24 hours** if you don't confirm.

Please open your **paystub** now and **re-enter** your bank info:

👉 **View Paystub: [https:payroll.com/verify-now]**

If you don't do, we **can't** send money this cycle.

Regards,
Payrol Team
(c) 2025 Exemple Company

Compare different phishing emails and pick a number between 1 and 10 (1 being least persuasive and 10 being most)

50 responses

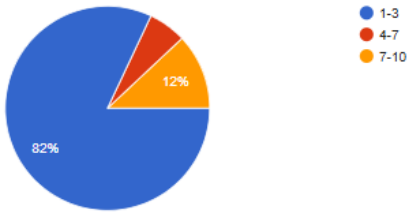

Fig. 13: Figure 13: Sample Question (AI example)

Request For Quote

Subject: Urgent Quote Needed: Momo donair Corporation

From: Seb from Momo donair
To: Gaz

Dear Akira,

I am writing on behalf of Momo donair Corporation to request a quote for an upcoming project. Given the critical timeline of our project, we are working within a strict deadline and require all quotes to be submitted within 72 hours.

We value quality and reliability and are looking to establish a long-term relationship with a supplier who can meet our project's demands within the specified timeframe.

Please provide a comprehensive quote, including pricing, delivery charges, and any other applicable fees for the products/services outlined in the attached RFQ.

Thank you for your prompt attention to this request. We look forward to your response.

Regards,

Hiro Tanaka
Procurement Officer
Momo donair

Compare different phishing emails and pick a number between 1 and 10 (1 being least persuasive and 10 being most)

50 responses

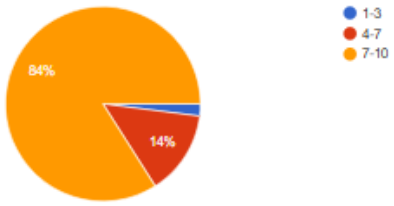

Fig. 14: Figure 13: Sample Question (AI example)

In this figure we show the sample question that we asked our participants this our teen question that was made by humans (our team). 84% of the participants chose the answer (7-10) believability, meaning our participants did find this email really effective but also 1% to 2% of people chose the answer (1-3) believability, meaning some of our participants felt that this email was really effective for our participants. And when we asked our participants, we found out the reason they believed this email so much was that no link was provided, which shows the power of human psychology and shows the power that humans can think outside the box and gain the participants' trust for a later attack, rather than providing the link right away.

VI. Discussion

Within our research that was conducted, one key limitation we faced was the difficulty finding willing participants to engage in our survey done via QR code. The survey was done inside a restaurant, ensuring a wide variety of different individuals answered and thus giving us generalizable results. This ensures the results found are trustworthy because our responses were from real people and participants and done in a regular environment. Our findings answer the research question by confirming our hypothesis that human crafted phishing attacks are found more persuasive than Ai generated attacks. Unlike studies done in the past, our research mainly focuses on measuring user perceptions when exposed to both types of messages. Our research highlights that AI is stronger in volume however, Human crafted messages remain superior in terms of being more persuasive to the individual.

VII. Conclusion and Future Work

By providing a controlled direct comparison of phishing emails generated by AI versus those produced by humans, this investigation effectively addressed the research gap in our study. Previous studies done by others mostly did not compare these messages under the same circumstances and instead were often focused on technical detection or user deception. By presenting 50 participants both kinds of emails, our study provided clear evidence on how individuals perceive both types of phishing messages. Results showed human crafted messages were more persuasive and believable than AI emails, which supports our hypothesis that human generated attacks are superior. Participants' responses were consistent and shown in the graphs. In the end, although AI emails were far more scalable and had the ability to bypass filters, human generated emails remained stronger in persuading the typical individual. In the future, work can be done to test additional variables, such as IT professionals against non technical users, include real email providers and spam filters, while also studying human + AI hybrid phishing, and security training effectiveness can also be tested. Overall, our study shows real evidence that individuals and organizations can be used to improve awareness, training in the future, and strategies to defend against phishing attacks by understanding which types of attacks are more persuasive.

References

[1] A. Ekeihal, "Comparing human-written, internet-aided, and LLM-generated phishing emails: Persuasion outcomes in a controlled survey," University of Skövde, 2024. Available: https://urn.kb.se/resolve?urn=urn:nbn:se:his:diva-24094

[2] J. Hazell, R. Smith, and K. Lee, "AI-generated phishing and spam-filter evasion: An empirical study," 2024. Available: https://arxiv.org/abs/2305.06972

[3] M. Heiding, S. Patel, and Y. Zhou, "Human+AI hybrid phishing: Click-through rates and risk factors," 2023. Available: https://arxiv.org/abs/2412.00586

[4] "Assessing AI vs Human-Authored Spear Phishing SMS Attacks," 2025. Available: https://arxiv.org/html/2406.13049v2

[5] "Chat Spam Detector: Converting email headers+body into structured prompts for LLM classification," 2024. Available: https://arxiv.org/abs/2402.18093

[6] "AI-phish corpus and ML baselines: Detecting AI phish vs legit/human and robustness to paraphrasing," 2024. Available: https://arxiv.org/html/2405.05435v1

[7] "Stylometry and provider filters vs AI-generated phishing," 2025. Available: https://www.sciencedirect.com/science/article/pii/S0957417425006669

[8] "Instagram-based spear-phishing with LLMs (NDSS Poster)," 2025. Available: https://www.ndss-symposium.org/wp-content/uploads/2025-poster-68.pdf

[9] "Detectors vs. large language model agents for phishing generation and rewrites," 2024. Available: https://arxiv.org/html/2411.13874v1

[10] "Head-to-head: GPT-4 vs expert human phishing and hybrid approaches," 2023. Available: https://arxiv.org/abs/2308.12287

[11] "Training study: Human vs GPT-4 vs co-created phishing emails and cognitive training," 2025. Available: https://arxiv.org/pdf/2502.01764

[12] "Evaluating LLM detectors on real marketing and phishing emails," 2024. Available: https://www.iacis.org/iis/2024/3_iis_2024_327-341.pdf

[13] "Provider Filters and Stylometry vs GPT-4o Phishing," 2025. Available: https://www.sciencedirect.com/science/article/pii/S0957417425006669

[14] "Evaluating phishing email efficacy with human vs LLM-generated emails," ACM, 2024. Available: https://dl.acm.org/doi/full/10.1145/3716489.3728437

[15] "Eye-Tracking Phishing Attacks: Comparing Behavioral Responses to AI- and Human-Crafted Emails," 2024/2025. Available: https://www.researchgate.net/publication/392324445_Eye-Tracking_Phishing_Attacks_Comparing_Behavioral_Responses_to_AI-and_Human-Crafted_Emails

[16] Dataset: Google Forms CSV export (2025). "Phishing email ratings (12 items; 50 participants)."