MDPI

Article

You Got Phished! Analyzing how to Provide Useful Feedback in Anti-Phishing Training with LLM Teacher Models

Tailia Malloy*[12], Laura Bernardy*[1], Omar El Bachyr[1], Fred Philippy[1], Jordan Samhi[1], Jacques Klein[1], Tegawendé F. Bissyandé[1]

[1]Interdisciplinary Centre for Security, Reliability and Trust (SnT) University of Luxembourg
[2] Zortify Labs, Zortify S.A. 19, rue du Laboratoire L-1911 Luxembourg
* Equal Contribution Corresponding Authors: tailia.malloy@uni.lu, laura.bernardy@uni.lu

Abstract

Training users to correctly identify potential security threats like social engineering attacks such as phishing emails is a crucial aspect of cybersecurity. One challenge in this training is providing useful educational feedback to maximize student learning outcomes. Large Language Models (LLMs) have recently been applied to wider and wider applications, including domain-specific education and training. These applications of LLMs have many benefits, such as cost and ease of access, but there are important potential biases and constraints within LLMs. These may make LLMs worse teachers for important and vulnerable subpopulations including the elderly and those with less technical knowledge. In this work we present a dataset of LLM embeddings of conversations between human students and LLM teachers in an anti-phishing setting. We apply these embeddings onto an analysis of human-LLM educational conversations to develop specific and actionable targets for LLM training, fine-tuning, and evaluation that can potentially improve the educational quality of LLM teachers and ameliorate potential biases that may disproportionally impact specific subpopulations. Specifically, we suggest that LLM teaching platforms either speak generally or mention specific quotations of emails depending on user demographics and behaviors, and to steer conversations away from an over focus on the current example.

Keywords: Cybersecurity, Phishing, Large Language Models, Education, Embeddings

1. Introduction

Recent advances in Generative Artificial Intelligence (GAI) including the advent of foundation models such as Large Language Models (LLMs) have been fundamentally transformative, demonstrating unprecedented performance across a wide range of tasks, including text generation, sentiment analysis, and question answering [1,2]. While the generalist nature of LLMs and other GAI models has facilitated their broad applicability, it poses significant limitations in scenarios requiring nuanced, user-specific responses [3], such as in educational contexts like anti-phishing training [4]. One of the most critical efforts to prevent social harm done by these new technologies is the effective training against social engineering, deepfakes of news, and other nefarious applications of GAI.

The complexity of social engineering attacks has significantly increased in recent months due in part to the advanced sophistication of GAI models [5]. These models can be used to quickly design new attacks from scratch using various methods such as translating previously used databases of attacks or creating complex novel attacks leveraging images, video, text, and audio [6] in an attempt to increase the success of social engineering attempts.

Received:
Revised:
Accepted:
Published:

Citation: . Analyzing Anti-Phishing Training. *Electronics* **2025**, *1*, 0. https://doi.org/

Despite the significant threat posed by GAI models such as LLMs in accelerating social engineering attacks [7–9], only 23% of companies polled by Proofpoint in 2024 had trained their employees on GAI safety [10].

One reason for this limitation of adequate training regarding GAI based phishing attacks is the high cost associated with traditional training methods such as in-person lecturing [11], and the time required to develop remote learning materials [12]. However, research has suggested that virtual learning of social engineering training can be more effective than in-person training [11]. A recent approach to addressing this educational limitation is to leverage GAI models themselves to design educational materials while providing feedback to users [4]. In the context of social engineering training in identifying phishing emails, this approach has the benefit of allowing for a training platform that can simultaneously generate realistic phishing attempts. While LLM supported training and education has benefits of easy access and scalability, it has issues related to domain specific knowledge and individualization of feedback in educational settings [13].

In this work we begin by presenting a dataset that serves to augment the original dataset presented by Malloy et al [4] containing a set of messages sent between human students and an LLM teacher in an anti-phishing education platform. We augment this dataset with two embedding dictionaries; the first is a set of embeddings of the messages sent by LLM teachers and human students; the second is a set of embeddings of open responses that students provided to describe the method that they used to determine if emails were phishing or not. This dataset includes embeddings formed by 10 different embedding models ranging from open to closed models and a range of embedding sizes.

After describing this presented dataset, we compare the 10 different embedding models in their correlation to human student learning outcomes. Next, we evaluate the usefulness of these embedding dictionaries by comparing the cosine similarity of the embeddings of messages sent by LLM teachers and students with the embeddings of the emails presented to students. These cosine similarity measurements are compared with several metrics of student learning performance, demographics, and other measures of the educational platform. We conclude this paper with a description of the results we present and a contextualization of these results with specific recommendations for improving LLM teaching methods.

2. Related Work

2.1. LLMs in Education

One example of a domain specific application of LLM education is discussed in [14] which focuses on databases and information systems in higher education. Here, the authors find that issues such as bias and hallucinations can be mitigated in domain specific educational applications through the use of an LLM-based chatbot 'MoodleBot', a specialized system tailored for a single specific educational course. These results highlight the importance of domain-specific knowledge in the design and evaluation of LLM teaching platforms. Meanwhile, a more generalist educational LLM platform is presented in [15] called multiple choice question generator (MCQGen), that can be applied to a variety of domains through the integration of Retrieval Augmented Generation and an human-in-the-loop process that ensures question validity.

Beyond applications of LLMs as merely educational tools is research into use cases of agentic LLMs that make decisions regarding student education. One recent survey by Chu et al. [16] of LLM applications on education focuses on the use of LLM agents, which extend the traditional use-case of LLMs beyond a tool into a more independent model that makes decisions and impacts an environment [17]. This survey highlights the importance of mitigating hallucinations and ensuring fairness in educational outcomes. This insight

guides an important focus of this work which compares different potentially vulnerable subpopulations in the way that they converse with an LLM chatbot. Many examples exist in the literature of LLM bias demonstrating potential causes of unfairness, such as racial bias [18], gender bias [19], or age [20]. These biases become increasingly relevant in domain specific applications of LLMs in education, as the ways in which biases interact with education become more complete than in other LLM applications [21].

2.2. LLM Personalization in Education

Personalization techniques have traditionally been extensively researched within information retrieval and recommendation systems but remain relatively underexplored in the context of LLMs [1]. Developing personalized and domain-specific educational LLMs involves leveraging user-specific data such as profiles, historical interactions, and preferences to tailor model outputs [22]. Effective personalization of LLMs is critical in domains such as conversational agents, education, healthcare, and content recommendation, where understanding individual preferences significantly enhances user satisfaction and engagement [22,23].

Recent literature highlights various strategies for personalizing LLMs, broadly categorized into fine-tuning approaches, retrieval augmentation, and prompt engineering [2,22,23]. Fine-tuning methods adapt LLM parameters directly to user-specific contexts, showing significant performance improvements in subjective tasks like sentiment and emotion recognition [2]. Fine-tuned LLMs have been applied onto educational domains such as the Tailor-Mind model which generates visualizations for use in educational contexts [24] However, these approaches are resource-intensive and often impractical for real-time personalization across numerous users [25].

Retrieval augmentation, on the other hand, enhances personalization efficiency by dynamically incorporating external user-specific information at inference time without extensive model retraining [26]. Methods like LaMP utilize user profiles and historical data, selectively integrating relevant context through retrieval techniques [1]. More recently, frameworks such as OPEN-RAG have significantly improved reasoning capabilities within retrieval-augmented systems, especially when combined with open-source LLMs [23]. Prompt engineering and context injection represent lighter-weight approaches where user-specific information is embedded within the prompt or input context, guiding the LLM toward personalized responses [22,27]. RAG has been applied on to domain-specific educational contexts like computing education [28] through the use of small LLMs that incorporate RAG. Other recent approaches in LLM education with RAG seek to personalize pedagogical content by predicting user learning styles [29], These methods, while efficient, are limited by context length constraints and impermanent personalization.

2.3. Automatic Phishing Detection

On the defensive side, research efforts are increasingly focused on countering these threats. The growing sophistication of LLM-generated phishing emails presents challenges for traditional phishing detection systems, many of which are no longer able to reliably identify such attacks. This issue has thus become a focal point in AI-driven cybersecurity research, which is particularly evident in the following two leading approaches.

[30] employed LLMs to rephrase phishing emails in order to augment existing phishing datasets, with the goal of improving the ability of detection systems to identify automatically generated phishing content. Their findings suggest that the detection of LLM-generated phishing emails often relies on different features and keywords than those used to identify traditional phishing emails.

LLM-generated phishing emails were also used in the approach of [31] to fine-tune various AI models, including BERT-, T5- and GPT-based architectures. Their results demonstrated a significant improvement in phishing detection performance across both human- and LLM-generated messages, compared to the baseline models.

2.4. LLM Generated Phishing Emails

Several studies have highlighted that generative AI can be leveraged to create highly convincing phishing emails, significantly reducing the human and financial resources typically required for the creation of them [30–35]. This development is driven in part by the increasing ability of LLMs to maintain syntactic and grammatical integrity while they also embed cultural knowledge into artificially generated messages [36]. Moreover, with the capacity to generate multimedia elements such as images and audio, GAI can enhance phishing emails by adding elements that further support social engineering attacks [32]. The collection of personal data for targeting specific individuals can also be facilitated through AI-based tools [35].

In [31], Bethany et al. evaluated the effectiveness of GPT-4-generated phishing emails and confirmed their persuasive power in controlled studies. A related study, revealed that while human-crafted phishing emails still demonstrated a higher success rate among test subjects, they were also more frequently flagged as spam compared to those generated by GPT-3 models.[33] Targeted phishing attacks—commonly known as spear phishing—can also be rapidly and extensively generated by low experienced actors using GAI, as demonstrated in [35] experiments with a LLaMA-based model.

2.5. Anti-Phishing Education

Anti-Phishing education seeks to train end-users to correctly identify phishing emails they receive in real life and react appropriately. This education is an important first step in cybersecurity as user interaction with emails and other forms of social engineering is often the easiest means for cyberattackers to gain access to privileged information and services [37]. Part of the ease with which attackers can leverage emails is due to the high number of emails that users receive as a part of their daily work, which leads to a limited amount of attention being placed on each email [38]. Additionally, phishing emails are relatively rare to receive as many filtering and spam detection methods prevent them from being sent to users' inboxes. For this reason, many users are relatively inexperienced with phishing emails and may incorrectly identify them [39]. Despite the commonality of cybersecurity education and training in many workplaces, social engineering including phishing emails remains a common method of attack with a significant impact on security [40].

Part of the challenge of anti-phishing education is defining the qualities of a good education platform and determining how to evaluate both the platform and the ability of users to detect emails in the real world. In their survey, Jampen et al. note the importance of anti-phishing education platforms that can equitably serve large and diverse populations in an inclusive manner [37]. This review compared 'user-specific properties and their impact on susceptibility to phishing attacks' to identify key features of users such as age, gender, email experience, confidence, and distrust. This is crucial as cybersecurity preparedness is only as effective as its weakest link, meaning anti-phishing education platforms that only work for some populations are insufficient to appropriately address the dangers associated with phishing emails [41]. It is important for anti-phishing training platforms to serve populations as they vary across these features to ensure that the general population is safe and secure from attacks using phishing emails [42].

Another important area of research uses laboratory and on-line experiments with human participants engaged in a simulation of anti-phishing training to compare different

approaches. This has the benefit of allowing for more theoretically justified comparisons, since traditional real-world anti-phishing education has high costs associated with it, making more direct comparisons difficult [40]. Some results within this area of research indicate that more detailed feedback, rather than only correct or incorrect information, significantly improves post-training accuracy in categorizing emails as either phishing or ham [43]. Additional studies indicate that personalizing LLM-generated detailed feedback to the individual user through prompt engineering can further improve the educational outcomes of these platforms [4,44]. However, these previous approaches do not involve the training or fine-tuning of more domain-specific models, and rely on off-the-shelf black box models using API calls to generate responses.

3. Dataset

3.1. Original Dataset

The experimental methods used to gather the dataset used for analysis in this work are described in [4] and made available on OSF by the original authors[1]. 417 participants made 60 total judgments about whether emails they were shown were safe or dangerous, with 8 different experiment conditions that varied the method of generating emails and the specifics of the LLM teacher prompting for educational feedback. These emails were gathered from a dataset of 1461 emails, with a variety of methods used to create these emails. In each of the four conditions we examine used educational example emails that were generated by a GPT-4 LLM model. While the experimentation methods contained 8 different conditions, we are interested only in the four conditions that involved conversations between users and the GPT-4.1 LLM chatbot. In each of these experiments, the participants were given feedback on the accuracy of their categorization from an LLM that they could also converse with.

Between these four conditions, the only difference was the presentation of emails to participants and the prompting of the LLM model for feedback. In the 'base' first condition, emails were selected randomly, and the LLM model was prompted to provide feedback based on the information in the email and the decision of the student. In the second condition, emails were selected by an IBL cognitive model in an attempt to give more challenging emails to the student, based on the past decisions they made. The third condition selected emails randomly but included information from the IBL cognitive model in the prompt to the LLM; specifically, this information was a prediction of which features of an email the current student may struggle with. Finally, the fourth condition combined the two previous ones, using the IBL cognitive model for both email selection and prompting.

In the original dataset, there are three sets of LLM embeddings of each email shown to participants using OpenAI API to access 3 embedding models ('text-embedding-3-large', 'text-embedding-3-small', and 'text-embedding-ada-002') [45]. These embeddings were used alongside a cognitive model of human learning and decision making called Instance Based Learning [46–48] to predict the training progress of users. However, the original paper [4] did not directly analyze the conversations between end users and the LLM chatbots, and did not create a database of chatbot conversation embeddings.

3.2. Proposed Dataset

In this work we introduce an embedding dictionary [2] of these messages and evaluate the usefulness of this embedding dictionary in different use cases. We also include in the same dataset an embedding dictionary of the open response replies that students gave at

[1] https://osf.io/wbg3r/
[2] https://osf.io/642zc/

the end of the experiment to answer the question of how they determined if emails were safe or dangerous. In the majority of our analysis we combine the four conditions that included conversations with chatbots because the previously mentioned differences do not impact conversations between participants and the LLM chatbot.

One major limitation to this previous dataset is the exclusive use of closed source models. While the embeddings themselves were included, the closed source nature of the three embedded models used in [4] limits the reproducibility of the work and the accessibility to other researchers. In this work we employ the same three closed source models as in the original work as well as seven new open source models (qwen3-embedding-0.6B[3] [49], qwen3-embedding-4B [4] [49], qwen3-embedding-8B [5] [49], all-MiniLM-L6-v2 [6], bge-large-en-v1.5 [7] [50], embeddinggemma-300m [8] [51], and granite-embedding-small-english-r2[9] [52]). For all models that did not directly output embeddings, mean pooling was used to extract embeddings [53]. There were 3846 messages sent between chatbots and 146 different users during the anti-phishing training, resulting in 38460 message embeddings in our dataset. Additionally, we provide embeddings for the seven new open source models of the emails in the original dataset resulting in 5856 new email embeddings.

Our conversation analysis presented in the following section begins by a comparison of the ten embedding models contained in our proposed dataset along a single metric. After this, we perform a series of regressions that compare correlations of different metrics of performance with the cosine similarity between the embeddings of messages and emails. Finally, we perform a mediation analysis to give more strength to our conclusions and recommendations. After this analysis we proceed to the Results and Discussion sections.

4. Conversation Analysis

In this section we demonstrate the usefulness of the presented dataset of embeddings between users and the teacher LLM in this anti-phishing education context. We begin by comparing the cosine similarity of the embeddings of messages sent by students and the LLM teacher with the emails that the student was viewing when the message was sent. This is an exploratory analysis that serves to examine whether cosine similarity is correlated with three different student performance metrics. An important aspect of this analysis is that it is purely correlational, meaning that we cannot determine causal relationships or the direction of correlational relationships. Our goal with this analysis is to explore potential methods of improving LLM education that can be further explored in future research. Code to generate all figures and statistical analysis in this section is included online[10].

4.1. Embedding Model Comparison

Before presenting our analysis of the correlations between cosine similarity and different attributes of student performance and demographics, we first seek to motivate our choice of cosine similarity as a metric. There are several more simple metrics that could be calculated between emails and messages without the need for embedding models, raising the question of the value of our proposed dataset. For instance, metrics of the lexical overlap between emails and messages such as the Jaccard [54], the proportion of common words between the message and the email, and the Rouge [55], a count of how many of the

[3] https://huggingface.co/Qwen/Qwen3-Embedding-0.6B
[4] https://huggingface.co/Qwen/Qwen3-Embedding-4B
[5] https://huggingface.co/Qwen/Qwen3-Embedding-8B
[6] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2
[7] https://huggingface.co/BAAI/bge-large-en-v1.5
[8] https://huggingface.co/google/embeddinggemma-300m
[9] https://huggingface.co/ibm-granite/granite-embedding-small-english-r2
[10] https://github.com/TailiaReganMalloy/PhishingConversations

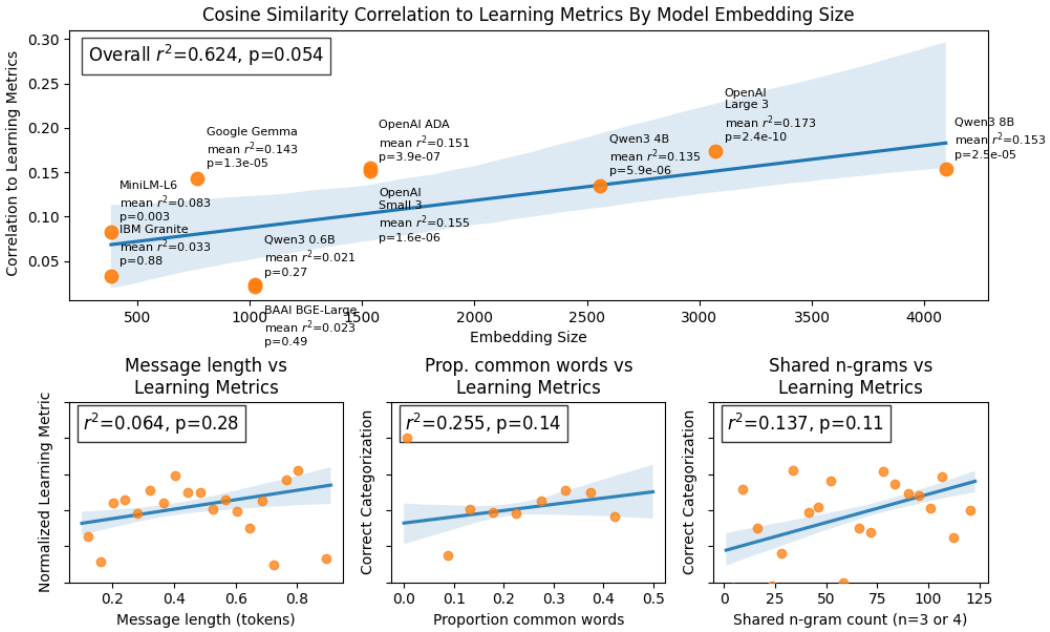

Figure 1. Top: A comparison of the correlation between three learning metrics and the cosine similarity of embeddings of messages sent by both teachers and students, and embeddings of email educational examples. This demonstrates significant correlation in the majority of models, and a general trend of increasing correlation with increasing embedding size. Bottom Left: The correlation between message length and the three learning metrics, demonstrating insignificant correlation. Bottom Middle: The correlation between the three learning metrics and the proportion of common words between emails and messages, also demonstrating insignificant correlation. Bottom Right: The correlation between the three learning metrics and the number of shared n-grams (2,3,4, or 5) between emails and messages, additionally demonstrating no significant correlation.

same short phrases (n-grams) appear in both texts. Additionally it is important to control for attributes such as the message length [56], since longer messages may have on average higher similarities to emails since they are of a similar length.

If the correlational analysis we present in this section could be equally related to these alternative metrics, it would demonstrate an issue with our proposition of the usefulness of the dataset we present. To address this, we begin by comparing the correlation of three metrics of student performance, their correct categorization, their confidence, and their reaction time. In Figure 1 we compare the correlation between three learning metrics and the cosine similarity of embeddings of messages sent as feedback and the emails students are observing. We report this average for 10 different embedding models. Additionally, we compare these correlations to the alternative metrics previously mentioned.

To determine which embedding model is ideal for our correlation analysis, we compare each of these embedding models in terms of the average correlation of our three metrics, correct categorization, confidence, and reaction time. This is shown on the top row of Figure 1 which has on the x axis the embedding size of the models under comparison, and on the y-axis the average correlation (Pearson R^2) of those three metrics. Overall we see a significant trend that models with larger embedding sizes are typically better correlated to the learning metrics we are interested in. This result is promising for our analysis, as comparing the similarity of larger embeddings often captures semantic similarity better than smaller embeddings [57]. The highest correlation to learning metrics is observed when comparing the cosine similarity of emails and messages generated by the Open AI Large 3 model which has an embedding size of 3048. For this reason we will compare the cosine similarity as measured using embeddings formed by the Open AI Large 3 model in all following analyses.

4.2. Regression Analyses

In all regression analyses in this section, we first bin the message embedding cosine similarities to email embeddings to the nearest 0.01, grouped based on the sender. Additionally, all message cosine similarity values on the x-axis are normalized to between 0-1 grouped by the message sender. This is the source of the values on the x-axis of each plot. Then, we plot as a scatterplot the averages for the metric on the y-axis of all of the binned messages. For example, in the left column of Figure 2, the leftmost blue point represents the average correct categorization for all trials where messages were sent that had embeddings with a cosine similarity to email embeddings of 0.00. The significance of these regressions is based on Pearson correlation coefficients with the R^2 and p values shown at the top of each subplot. Finally, each variable comparison (e.g correct categorization and message cosine similarity to email) has a T-Test run to compare the correlation in a different manner that does not use binned message cosine similarity values.

The first of these metrics is the percent of correct categorization by the student, the second is their confidence in the categorization, and the last is the reaction time of the student. Ideally, the teacher LLM would be providing feedback that is easy to quickly understand and leads to high confidence and correct categorizations. These three metrics are compared to the cosine similarity of emails with respect to both student and teacher message embeddings as shown in Figure 2.

4.3. Categorization Accuracy

The relationship between message cosine similarity and user categorization accuracy is shown on the middle column of Figure 2. The analysis of student accuracy in categorization revealed that both the human student's and teacher LLM message cosine similarities to emails were positively associated with the likelihood of a correct categorization. The human student's message-email cosine similarity showed a moderate positive correlation with correct categorization, that is not robust when evaluated with ANOVA (Pearson Correlation: $R^2 = 0.243, p = 0.0197$, ANOVA: $F(22, 464) = 0.841$, $p = 0.674$, $\eta_p^2 = 0.038$). The teacher LLM's message similarity exhibited a strong positive association with correct outcomes, a relationship further corroborated by statistically significant results from ANOVA, though the effect size was small (Pearson Correlation $R^2 = 0.578, p = 6.66x10^{-6}$, ANOVA: $F(25, 1720) = 1.648$, $p = 0.0231$, $\eta_p^2 = 0.023$) These results indicate that student performance was higher when the messages sent by either them or their teacher were more closely related to the email that was being observed by the student. Furthermore, the results suggest that the LLM's message similarity is a stronger predictor of correct categorization than the human student's similarity, though the ANOVA effect sizes remain modest.

4.4. Categorization Confidence

The relationship between message cosine similarity and user confidence in their categorization is shown in the middle column of Figure 2. The analysis of students' categorization confidence showed a divergent trend for the student and the teacher in relation to message similarity. This is a surprising result, since the previous analysis of categorization accuracy indicated that both student and teacher messages that were more related to the current email were associated with better performance. However, confidence is a separate dimension from accuracy as low confidence correct answers and high confidence incorrect answers can change the relationship between message embedding similarities and this metric of student performance. The cosine similarity between a student's message and the email content was negatively associated with the student's confidence rating (Pearson Correlation: $R^2 = 0.269, p = 0.0133$, ANOVA: $F(22, 464) = 1.539$, $p = 0.0569$, $\eta_p^2 = 0.068$).

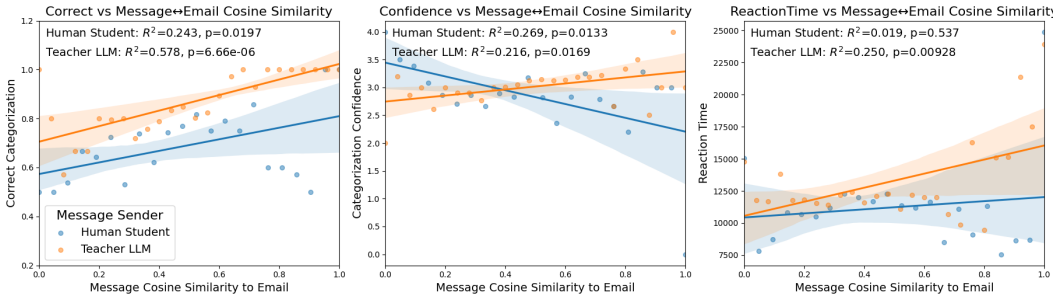

Figure 2. On all plots, orange indicates messages sent by the teacher LLM and blue represents messages sent by the human student. Shaded regions represent 95% confidence interval and similarities are binned to the nearest 0.01. Left: A correlation analysis between message cosine similarity to emails with the probability of correct categorization, showing significant correlation for both types of messages. Middle: A correlation analysis of message cosine similarity to emails and student confidence in their categorization, indicating a significant positive correlation for teacher messages and a significant negative correlation for student messages. Right: A correlation analysis between student reaction time and message cosine similarity to emails, indicating no significance for messages sent by students but a significant, but moderate, positive trend for messages sent by teachers.

In other words, students who more closely echoed the email's content in their own messages tended to report lower confidence in their categorization decisions, but this pattern was not consistently supported across groups, as indicated by ANOVA score. By contrast, the teacher LLM's message similarity showed a positive correlation with student confidence which was also statistically significant in ANOVA (Pearson Correlation: $R^2 = 0.216, p = 0.0169$, ANOVA: $F(25, 1720) = 1.652$, $p = 0.0225$, $\eta_p^2 = 0.023$). This indicates that when the teacher's response closely matched the email content, students tended to feel slightly more confident about their categorizations, although the effect size was small.

4.5. Categorization Reaction Time

The relationship between message cosine similarity and reaction time is shown on the right hand side of Figure 2. The relationship between reaction time and message similarity differed markedly by role. There was no significant association between the human student's message similarity and their reaction time, not with Pearson Correlation nor with ANOVA (Pearson Correlation: $R^2 = 0.019, p = 0.537$, ANOVA: $F(22, 464) = 1.155$, $p = 0.284$, $\eta_p^2 = 0.052$), indicating that how closely a student's message mirrored the email content did not measurably influence how quickly they responded. In contrast, the teacher LLM's message similarity was significantly associated with longer reaction times in regard to Pearson Correlation, but ANOVA also showed just a small effect size (Pearson Correlation: $R^2 = 0.250, p = 0.0093$, ANOVA: $F(25, 1720) = 0.882$, $p = 0.632$, $\eta_p^2 = 0.013$).

Higher cosine similarity between the teacher's message and the email corresponded to increased time taken by students to complete the categorization task, even if the effect was not conventionally significant with ANOVA, it shows a trend. In practical terms, when the teacher's response closely resembled the email text, students tended to require more time to finalize their categorization, whereas the student's own content overlap had little to no observable effect on timing. These results are presented as correlational patterns (from the regression analysis) and do not imply causation, but they highlight that teacher-provided content overlap was linked to slower student responses while student-provided overlap was not.

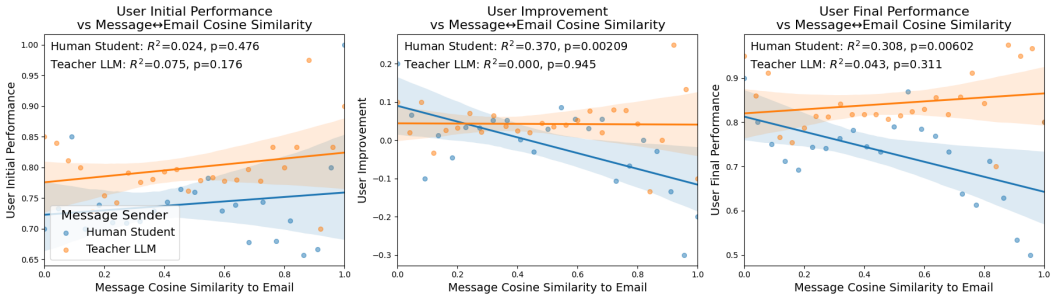

Figure 3. On all plots, orange indicates messages sent by the teacher LLM and blue represents messages sent by the human student. Shaded regions represent 95% confidence interval and similarities are binned to the nearest 0.01. Left: A correlation analysis between message cosine similarity to emails with user initial performance, showing no significant correlation for both types of messages. Middle: A correlation analysis of message cosine similarity to emails and student categorization improvement, indicating a significant negative correlation for student messages and a no significant correlation for teacher messages. Right: A correlation analysis between student final performance and message cosine similarity to emails, indicating no significance for messages sent by teachers but a significant negative trend for messages sent by students.

4.6. Student Learning Outcomes

The next analysis that we perform is related to the learning outcomes of the students, as well as their responses to the post-experiment questionnaire that asked them questions about whether they thought the emails that they observed were written by humans or an LLM. Note that in the three conditions we examine here, all of the emails were written and stylized with HTML and CSS code by a GPT-4.1 LLM, meaning that the correct perception of emails as AI generated is 100 percent. The open response question that is analyzed on the right column of figure 3 is the student's response to the question of how they made their decisions about whether an email was safe or dangerous.

4.6.1. User Initial Performance

The left column of figure 3 compares the average message cosine similarity to the current email being observed by the student with the initial performance of the student. Here we see that neither the messages sent by human students nor the teacher LLM are strongly correlated with user initial performance. There is a slight positive trend for both regressions where higher cosine similarity with student messages is associated with better initial performance (Pearson Correlation: $R^2 = 0.024, p = 0.476$, ANOVA: $F(22, 464) = 0.692$, $p = 0.849$, $\eta_p^2 = 0.032$), and similarly for teacher LLM similarity (Pearson Correlation $R^2 = 0.075, p = 0.176$, ANOVA: $F(25, 1720) = 0.863$, $p = 0.659$, $\eta_p^2 = 0.012$). However, both of these have low correlations with high p-values and the ANOVA results show no significance and small effect sizes. This indicates that there is no relationship between the conversations of human students and LLM teachers and initial performance, at least when measured by message cosine similarity to emails. This makes intuitive sense as the messages between participants and students begin after this initial pre-training phase when there is no feedback yet.

4.6.2. User Training Outcomes

The middle column of figure 3 compares the user improvement to our measure of message cosine similarity to emails. Here, we can see that only the messages sent by human students have cosine similarities to emails that are correlated with user improvement, supported by Pearson Correlation. However, interestingly this is actually a negative trend, meaning that higher human message cosine similarity to emails results in lower average user improvement (Pearson Correlation: $R^2 = 0.370, p = 0.002$ ANOVA: $F(22, 464) = 1.557$,

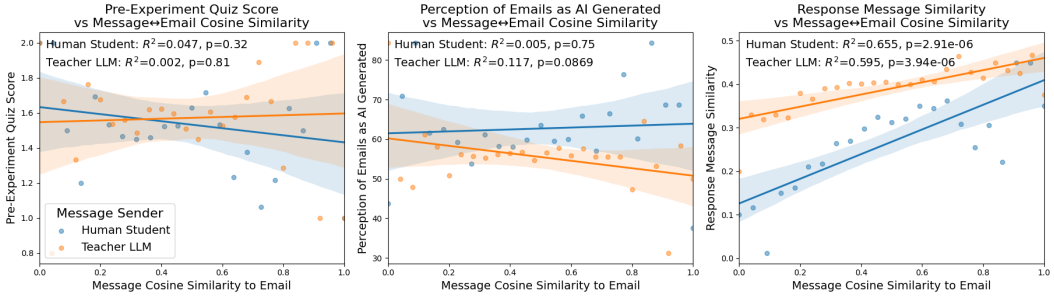

Figure 4. On all plots, orange indicates messages sent by the teacher LLM and blue represents messages sent by the human student. Shaded regions represent 95% confidence interval and similarities are binned to the nearest 0.01. Left: A correlation analysis between message cosine similarity to emails with user initial performance, showing no significant correlation for both types of messages. Middle: A correlation analysis of message cosine similarity to emails and student categorization improvement, indicating a significant negative correlation for student messages and a no significant correlation for teacher messages. Right: A correlation analysis between student final performance and message cosine similarity to emails, indicating no significance for messages sent by teachers but a significant negative trend for messages sent by students.

$p = 0.0521, \eta_p^2 = 0.069$)), though the ANOVA result is not statistically significant, indicating that the effect is not robust across groups. Meanwhile, this same comparison of teacher LLM messages shows no correlation at all (Pearson Correlation: $R^2 = 0.000, p = 0.945$ ANOVA: $F(25, 1720) = 1.014$, $p = 0.444$, $\eta_p^2 = 0.015$). This goes against the intuition that conversations that focus on the content of emails are beneficial to student learning outcomes that were established in the previous set of results. However, we believe they are not completely contradictory as a human student sending messages about specific parts of emails, even including specific passages of the email, may indicate a high level of confusion about the categorization.

4.6.3. User Final Performance

The right column of figure 3 compares the user improvement to the message cosine similarity to emails. Similarly to the comparison to user improvement, here we see no correlation with the LLM teacher messages and user final performance (Pearson Correlation $R^2 = 0.043, p = 0.311$ ANOVA: $F(25, 1720) = 1.189$, $p = 0.237$, $\eta_p^2 = 0.017$), while the human emails have a similar negative correlation (Pearson Correlation $R^2 = 0.308, p = 0.006$ ANOVA: $F(22, 464) = 1.705$, $p = 0.0247$, $\eta_p^2 = 0.075$). Both correlation measures support these outcomes. This supports the conclusions of the previous comparison of regressions which suggested that participants who frequently make comments that reference specific parts of the emails they are shown may have worse training outcomes. Taking these results in mind while observing the results of regressions shown in Figure 2 suggests that LLM models should seek to make their feedback specific and reference the emails that are being shown to participants, but steer human participants away from focusing too much on the specifics of the email in question in their own messages.

4.7. Student Quiz Responses

The next set of cosine similarity analyses that we perform using the cosine similarity of messages and emails compares the performance of students on the quizzes they completed before and after training.

4.7.1. Student Pre-Experiment Quiz

The left column of Figure 4 compares the pre-experiment quiz score of students to the message cosine similarity between the emails and the messages sent by human

students and LLM teachers. Here we see no correlation between the messages sent by either students (Pearson Correlation: $R^2 = 0.047, p = 0.32$, ANOVA: $F(22, 464) = 1.195$, $p = 0.247$, $\eta_p^2 = 0.054$) or teachers (Pearson Correlation: $R^2 = 0.002, p = 0.81$, ANOVA: $F(25, 1720) = 1.261$, $p = 0.174$, $\eta_p^2 = 0.018$). As with the user initial performance, this makes intuitive sense since the base level of student ability shouldn't have a direct impact on the way that students and teachers communicate relative to the email that the student is observing. One potential difference between these communications that is not directly measured in this analysis is the information within the email itself that may be focused on more or less in conversations depending on student initial ability.

4.7.2. Student Post-Experiment Quiz

The middle column of Figure 4 compares user participant perception of emails as being AI generated and the similarity of messages sent between human students and LLM teachers and the current email being observed. Here we see no correlation for messages sent by human students (Pearson Correlation: $R^2 = 0.005, p = 0.75$, ANOVA: $F(22, 464) = 1.348$, $p = 0.135$, $\eta_p^2 = 0.060$) or for messages sent by the LLM teacher (Pearson Correlation $R^2 = 0.117, p = 0.869$, ANOVA: $F(25, 1720) = 0.702$, $p = 0.86$, $\eta_p^2 = 0.010$). There is a slight negative trend here observable as pattern, where a lower perception of emails as being AI generated is slightly associated with a lower LLM teacher message cosine similarity. This is an interesting trend as the true correct percentage of emails that are AI generated is 100%, however this trend is statistically not significant.

4.7.3. Student Post-Experiment Open Response

The right column of Figure 4 compares the similarity between the current email being observed by a student and the open response messages that they gave to the question of how they made their decisions of whether emails were safe or dangerous. Here we see the strongest and most significant trend over all of the embedding similarity regressions we have performed. There is a strong positive trend for human student messages with both correlation measures (Pearson Correlation: $R^2 = 0.655, p < 1e-3$, ANOVA: $F(22, 464) = 5.624$, $p = 4.86e-14$, $\eta_p^2 = 0.211$) and LLM teacher messages (Pearson Correlation $R^2 = 0.595, p < 1e-3$, ANOVA: $F(25, 1720) = 1.377$, $p = 0.102$, $\eta_p^2 = 0.020$) where the more similar a message is to the email that the human student is observing, the more similar that message is to the open response question at the end of the experiment. For the LLM teacher messages, this effect shows less robust according to ANOVA.

4.8. User Demographics

The final set of cosine similarity regressions we perform compares the similarity of messages sent by human students and LLM teachers and the different demographics measurements that were included in the original dataset.

4.8.1. Age

Comparing the age of participants and their conversations demonstrates a significant correlation to the messages sent by human students (Pearson Correlation: $R^2 = 0.315, p = 0.005$, ANOVA: $F(22, 464) = 1.395$, $p = 0.11$, $\eta_p^2 = 0.062$), and an insignificant but present trend for the messages sent by the Teacher LLM (Pearson Correlation: $R^2 = 0.115, p = 0.0904$, ANOVA: $F(25, 1720) = 1.122$, $p = 0.307$, $\eta_p^2 = 0.016$). Both of these correlations trend negative, indicating that older participants have less correlation in the messages they send and the emails they are currently observing. ANOVA confirms a small-to-moderate effect size of this for the student-message similarity over groups, while being not conventionally significant, since it can't be consistently observed across all groups.

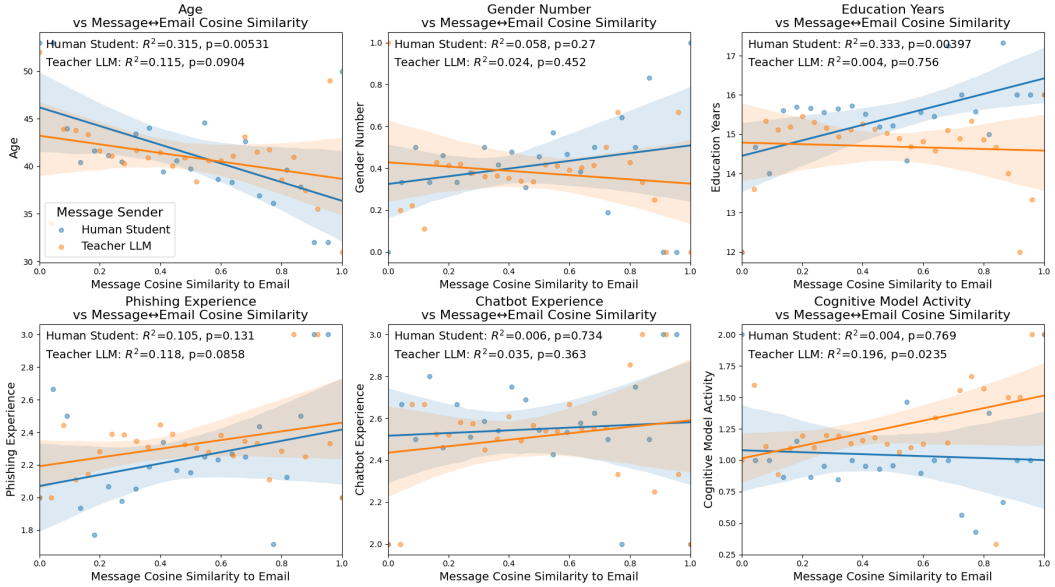

Figure 5. On all plots, orange indicates messages sent by the teacher LLM and blue represents messages sent by the human student. Shaded regions represent 95% confidence interval and similarities are binned to the nearest 0.01. Top Left: A correlation analysis between message cosine similarity to emails with user age, showing a significant negative correlation for both types of messages. Top Middle: A correlation analysis of message cosine similarity to emails and student gender, indicating no significant correlation for student messages and a no significant correlation for teacher messages. Top Right: A correlation analysis between student number of years of education and message cosine similarity to emails, indicating no significance for messages sent by teachers but a significant positive trend for messages sent by students. Bottom Left: A correlation analysis between message cosine similarity to emails and phishing experience, as measured by a pre-experiment questionnaire, indicating a insignificant positive trend for both types of messages. Bottom Middle: A correlation analysis of experience with chatbots, as measured by a pre-experiment questionnaire, and message cosine similarity to emails, indicating no significant trend for either type of message. Bottom Right: A correlation analysis comparing message cosine similarity to emails and cognitive model activity, as measured by the condition of the experiment, indicating a significant positive trend for teacher messages and no trend for student messages.

4.8.2. Gender

To perform a regression in the same format as the previous analyses, we arbitrarily assigned female to a value of 1 and male to a value of 0 (there were 0 non-binary students in this subset of the original dataset). This allowed for an analysis, shown in the top-middle of Figure 5, which shows no correlation between the gender number of students and the messages sent by either human students (Pearson Correlation: $R^2 = 0.058, p = 0.27$, ANOVA: $F(22, 464) = 1.110$, $p = 0.331$, $\eta_p^2 = 0.050$) or by teacher LLMs (Pearson Correlation: $R^2 = 0.025, p - 0.452$), ANOVA: $F(25, 1720) = 0.880$, $p = 0.635$, $\eta_p^2 = 0.013$). This indicates that male and female students sent similar messages, and that the LLM replied with similar messages. While these results are insignificant, they do suggest that accounting for gender differences in how LLM teaching models give feedback to students is less of a priority compared to other subpopulations of students.

4.8.3. Education

Comparing the similarity of embeddings of messages sent between human students and LLM teachers demonstrates a correlation with the years of education that the student has received for messages sent by the human student (Pearson Correlation: $R^2 = 0.33, p = 0.003$), ANOVA: $F(22, 464) = 0.991$, $p = 0.474$, $\eta_p^2 = 0.045$) but not for the messages sent

by the teacher LLM ($R^2 = 0.004, p = 0.756$, ANOVA: $F(25, 1720) = 0.984$, $p = 0.486$, $\eta_p^2 = 0.014$). The positive trend between the number of years of education and the human student message cosine similarity to emails indicates that students with higher education send messages that more closely match the information contained in the emails they are observing. This effect is continuous but not significant with ANOVA, so it is more a trend showing than a stepwise jump between education categories. As mentioned with regards to age, education level is another important group to account for when improving educational outcomes, meaning education level could be a target for future improvement in LLM teacher feedback.

4.8.4. Phishing Experience

The next analysis we performed compared the level of phishing experience of human students, as measured by the response that students gave to the number of times that they have received a phishing email. We again mapped this discrete categorization onto a value to perform a regression. When we compare this measure of experience to the cosine similarity of messages sent and emails, we see no significant correlation in either messages sent by human students (Pearson Correlation: $R^2 = 0.105, p = 0.131$, ANOVA: $F(22, 464) = 0.923$, $p = 0.565$, $\eta_p^2 = 0.042$) or the teacher LLM (Pearson Correlation: $R^2 = 0.118, 0.085$, ANOVA: $F(25, 1720) = 0.912$, $p = 0.589$, $\eta_p^2 = 0.013$). While insignificant, both of these regressions demonstrate a slightly positive trend suggesting that more experienced users may be more likely to send messages related to the emails they are observing.

4.8.5. Chatbot Experience

Similar to phishing experience, chatbot experience was determined by mapping a multiple choice question onto values to allow for a regression. Interestingly, we see no correlation between email embeddings and the embeddings of messages sent by either human students (Pearson Correlation: $R^2 = 0.006, p = 0.734$, ANOVA: $F(22, 464) = 1.332$, $p = 0.144$, $\eta_p^2 = 0.059$) or teacher LLMs (Pearson Correlation: $R^0.035, p = 0.363$, ANOVA: $F(25, 1720) = 1.016$, $p = 0.442$, $\eta_p^2 = 0.015$), with both regressions displaying near 0 trends and high p-values. This indicates that the conversations during training were equally likely to be related to the emails that were being observed by participants whether the student had little or a high amount of experience with LLM chatbots. Typically we would assume that participants would converse differently if they had more experience, but here it is important to note we are comparing one specific aspect of the messages, whether they are related to the email being observed, meaning other comparisons of these conversations may display a difference across chatbot experience level.

4.8.6. Cognitive Model Activity

The final regression that we perform looked at the 'cognitive model activity', which is a stand-in for the condition of the experiment. While not directly a demographic, this did compare the messages sent by humans and the LLM based on the condition of the experiment. This metric was determined based on whether the IBL cognitive model used in the experiment performed no role (0), either determined the emails to send to participants or was used to prompt the LLM (1), or if the IBL model performed both of these tasks (2).

Comparing this measure of cognitive model activity which differed across experiment conditions demonstrates a positive and significant trend for messages sent by the LLM teacher, though the ANOVA shows no significant group-level effect (Pearson Correlation: $R^2 = 0.196, p = 0.023$, ANOVA: $F(25, 1720) = 1.159$, $p = 0.267$, $\eta_p^2 = 0.017$). This indicates that LLM messages are more likely to align with emails when the cognitive model is more active, even if differences across groups are minimal. For human student messages, the Pearson correlation shows no significant relationship, but ANOVA indicates signifi-

cant differences across conditions (Pearson Correlation: $R^2 = 0.004, p = 0.769$, ANOVA: $F(22, 464) = 1.725$, $p = 0.0222$, $\eta_p^2 = 0.076$). This suggests that while overall message similarity is not linearly correlated with cognitive model activity, there are measurable differences in how students respond depending on the experimental condition.

Table 1. Significant Mediation Effects on Correct Categorization

Context	Ind Coef.	SE	p	CI 2.5%	CI 97.5%	Sig
(Student+Teacher) ∼ Age	-0.00586	0.00238	0.012	-0.0109	-0.0019	Yes
(Student+Teacher) ∼ AI Gen Perception	0.00395	0.00220	0.044	0.000156	0.00834	Yes
(Student+Teacher) ∼ Response Msg Similarity	0.00884	0.00347	0.004	0.00285	0.0157	Yes
(Teacher) ∼ Education Years	-0.00582	0.00292	0.020	-0.0131	-0.00137	Yes
(Teacher) ∼ Response Msg Similarity	0.00922	0.00296	0.000	0.00459	0.0165	Yes

4.9. Mediation Analysis

In addition to regression analyses, we are interested in the impact of each of the demographic variables (Figures 3-4) on measures of student performance (Figure 1-2). For brevity, we focus our analysis on demographic measure impact on correct categorization. Mediation analysis can be used to test whether the impact of a variable X on Y is at lest partially explained by the effects of an intermediate variable M, called the 'mediator' [58]. This analysis is commonly used in social psychology [59], human-computer interaction [60], and in LLM research such as investigations of potential gender biases in LLMs [61]. This analysis was performed using the Pingouin python library package [62]. All of the significant mediation effects reported in this section are summarized in Table 1.

4.9.1. Mediation of Student and Teacher Messages

Our first set of mediation analyses compared whether the impact of demographic variables on student correct categorization could be mediated by the cosine similarity of student or teacher message embeddings and email embeddings. Out of all of the demographic variables, three significant mediation effects were observed. The first was Age, which had a significant total effect $(coef = 0.046, SE = 0.0211, p = 0.0299, CI95\%[0.00448, 0.0874])$, a significant direct effect $(coef = 0.0518, SE = 0.0211, p = 0.0141, CI95\%[0.0104, 0.0932])$, and a significant indirect effect $(coef = -0.00586, SE = 0.00238, p = 0.012, CI95\%[-0.0109, -0.0019])$.

The next significant mediation effect was of the effect of AI generation perception on correct categorization. There was a significant total effect $(coef = -0.12, SE = 0.021, p = 1.293e-08, CI95\%[-0.161, -0.0788])$, direct effect $(coef = -0.124, SE = 0.0209, p = 3.741e-09, CI95\%[-0.165, -0.0829])$ and indirect effect $(coef = -0.00395, SE = 0.0022, p = 0.044, CI95\%[0.000156, 0.00834])$.

The final significant mediation effect when using both student and teacher message similarities as a mediator is Response Message Similarity which had a significant total effect $(coef = 0.235, SE = 0.0206, p = 2.062e-29, CI95\%[0.195, 0.275])$, direct effect $(coef = 0.226, SE = 0.0208, p = 6.212e-27, CI95\%[0.185, 0.267])$, and indirect effect $(coef = 0.00884, SE = 0.00347, p = 004, CI95\%[0.00285, 0.0157])$

4.9.2. Mediation of Teacher Messages Only

The first of the two significant mediation effects with respect to teacher message cosine similarity to emails is on Education Years which had a significant total effect ($coef = -0.109, SE = 0.0238, p = 5.421e-06, CI95\%[-0.155, -0.0619]$), direct effect ($coef = -0.103, SE = 0.0237, p = 1.557e-05, CI95\%[-0.149, -0.0562]$), and indirect effect ($coef = -0.00582, SE = 0.00292, p = 0.02, CI95\%[-0.0131, -0.00137]$).

The second significant effect of teacher messages is that of Response Message Similarity which had a significant total effect ($coef = 0.217, SE = 0.0234, p = 4.474e-20, CI95\%[0.171, 0.263]$), direct effect ($coef = 0.208, SE = 0.0234, p = 1.665e-18$ $CI95\%[0.162, 0.254]$), and indirect effect ($coef = 0.00922, SE = 0.00296, p = 0, CI95\%$ $[0.00459, 0.0165]$).

5. Results

Before beginning our summary and interpretation of the above analysis, it is important to reiterate the reason for our analysis and the meaning of statistical significance in the context of correlational analysis. We initially sought to perform an exploratory analysis to compare potential areas of improvement in LLM teacher feedback generation. This is a broad question with many possible methods of improvement, necessitating the narrowing down of potential methods. While these results point to possible methods of improvement, they do not exclude the possibility that alternatives may also lead to improvements in LLM teacher feedback.

Across the 30 regressions that we performed, 12 reached statistical significance and 6 of the others showed meaningful trends. We additionally performed 25 mediation analyses which showed 5 significant mediation effects. Taken together, these results form a coherent picture of how email embedding cosine similarity to embeddings of messages sent by both students and teachers relates to student performance and learning. The following analyses summarize and synthesize the results of our message-email similarity analysis and make actionable recommendations for both real-world online training platforms and future studies of human learning using natural language feedback provided by LLMs.

In categorization accuracy, both student and teacher message-email similarity were positively correlated with student categorization accuracy. The similarity between the student message and email with the emails showed a moderate effect that is not consistent across categories, while this same metric for the LLM teacher messages showed a strong effect that is also confirmed by ANOVA score, suggesting that the LLM teacher messages that closely aligned with the observed email were most useful for guiding correct responses. This makes intuitive sense since feedback for participants that references the email they are currently observing would typically be more relevant than less email-related feedback. While this analysis is correlational, it does demonstrate one area that future LLMs could be trained to optimize, by encouraging or preferring responses that are more closely related to the emails that students are currently categorizing.

Building on this, the confidence results gave important insights: students who echoed the content of the emails more closely actually felt less confident, whereas higher teacher message-email similarity increased confidence. This correlation indicates that a student who is frequently making questions or comments that directly reference parts of the emails in the training might indicate someone who needs more experience and varied feedback to achieve higher improvement levels. Taken together, these findings indicate that while alignment with the email improves accuracy, when it is the student providing the overlap, it may signal uncertainty, whereas teacher-provided overlap reassures students.

For LLM-supported learning platforms, this finding suggests potential avenues for development toward greater user engagement in the conversation with the LLM teacher

during email categorization, which also extends the insights reported in [4,44]. Also, such a level of guidance cannot be provided in traditional in-person training [11].

With taking also reaction times into account, we found that teacher message-email similarity was significantly associated with longer response times in terms of Pearson Correlation, while student message cosine message-email similarity showed no significant effect. The ANOVA results suggested that neither variable exerted a statistically significant effect. While accuracy and confidence are clear objectives for improvement, which are both increased with higher teacher message-email similarity, a preference for either higher or lower reaction time is less obvious. Taking these results on reaction into account with the previous correlation analysis may indicate that while teacher message-email similarity may improve the important metrics of accuracy and confidence, that may come at the cost of a longer time requirement for students. In some educational scenarios this may be a trade-off, if student time is a significantly constrained resource. However, in other settings the improvement on accuracy and confidence correlated with teacher message-email similarity may be much more important, meaning the increased time requirement is relatively irrelevant.

When we turn from immediate task performance to learning outcomes, the results show a different pattern. Greater similarity between student message and the email text was negatively associated with both learning improvement and final performance, while teacher message-email similarity showed no significant relationship in either case. So high student message-email similarity predicts weaker learning, suggesting that anti-phishing training should encourage flexible strategies rather than focusing on specific examples. Taken together with the results on confidence and accuracy, this set of findings indicates that student message-email similarity is positively associated with immediate correctness but negatively associated with learning gains and final outcomes, while teacher message-email similarity is linked to immediate performance benefits without clear effects on longer-term improvement.

Students who closely echo phishing emails may rely too heavily on surface features, indicating that training should emphasize broader pattern recognition rather than simple repetition. Overall, over-reliance on specific email features may hinder broader learning and decision-making, highlighting the importance of teaching generalizable strategies for identifying phishing attempts. This becomes particularly important in the context of GAI-generated phishing emails, as these may be detected based on patterns beyond mere textual features [31–33].

The strongest effects we observed came in the post-experiment open responses, where both student and teacher message-email similarity were strongly and positively related to the strategies that students reported using. This is an interesting result as the open response questions ask the student to reply on their general strategy, rather than a specific email they observed. This indicates that it may be useful for LLM teachers to discuss the strategies that students use to determine if an email is safe during their feedback conversations with students.

Because these questions asked students to describe their general strategy rather than respond to a specific email, this suggests that anti-phishing training could benefit from emphasizing strategy development over rote memorization of individual examples [43], building on what was previously outlined in this regard. GAI LLMs could support this by providing strategy-focused feedback and offering personalized prompts when students rely too heavily on surface cues, as well as guiding post-task reflection on the strategies applied. Together, these approaches could help address a common challenge in phishing defense: students' tendency to over-rely on specific email features rather than developing robust, transferable detection strategies.

Our analysis of the LLM teacher and human student message-email similarity with respect to student demographics revealed important implications for improving diversity, equity, and inclusion in online training platforms. One of the most important aspects of equality in education is effective progress for students of all ages. This is especially important in anti-phishing education and the elderly as they are one of the most susceptible subpopulations with regards to phishing attempts [63]. These results indicate that older age groups may be less likely to have conversations about the specific emails they are observing in anti-phishing training. Taking this into account when providing natural language educational feedback could improve the learning outcomes of more aged individuals.

Finally, our analysis of mediation effects provided further support to the evidence that the types of messages that are sent between students and teachers, as well as the types of messages teachers send independently, can alter the impact of demographics on student performance. This was identified by the five mediation analyses with significant indirect effects, demonstrating that part of the impact of these demographics effects on correct categorization can be explained in part by effects related to how students and teachers communicate. This indicates that by improving the way that teacher LLMs, such as by using the metrics we suggest in this section, there maybe a similar improvement in the performance of students that reduces the biases related to specific subpopulations of students. This is a major target for improving LLM teacher quality, as it can potentially lead to more equitable outcomes in the application of LLM teachers, and reduce some of the concern over their widespread adoption.

6. Discussion

In this work we present a dataset of embeddings of messages sent between LLM teachers and human students in an online anti-phishing educational platform. The goal of this dataset is to be applied onto improving the quality of LLM teacher educational feedback in a way that can account for potential biases that exist within LLMs that raise concerns regarding their widespread adoption. Our analysis revealed relationships between metrics of educational outcomes and the semantic alignment of educational feedback discussions, as measured by the cosine similarity of message embeddings and the educational email embeddings. In general, we found that when the LLM teacher's feedback closely mirrored the content of the email under discussion, students performed better on the immediate task. We additionally found some correlations between these educational outcomes and the similarity of student messages to email examples, but overall the conclusions were more mixed compared to the analysis of teacher messages. Additionally, our mediation analysis provided further support that teacher message and email embedding similarity can serve as a mediator for the effect of several important demographics on the impact of student performance.

These results suggest that message-email similarity can be an important target for testing methods in training, fine-tuning, and prompting without the requirement of running additional tests with human subjects which can be costly, or relying on simulated LLM students which can have issues transferring to real world student educational improvement. Moreover, these results have applications outside of describing targets for testing methods by detailing some of the most important subpopulations to focus on for improvement of the quality of LLM teacher responses in the content of anti-phishing training. Specifically, age, education, phishing experience and experience with AI chatbots were identified as demographics in which certain subpopulations may be disproportionally negatively impacted by lower quality teacher LLMs. Our mediation analysis, as well as ANOVA and regression analyses, provided evidence that improving the quality of LLM teacher

responses using the methods we suggest can have a positive impact on the educational outcomes of these subpopulations.

Another possible approach to incorporate the lessons learned from this work into the design of new LLM teaching models is to attempt to detect and address learner confusion over phishing emails proactively. The negative correlation of student message similarity with learning outcomes indicates that over-fixation on specific aspects of the email examples can be a real-time signal that LLM teachers can use to adjust their feedback. Whether done through chain of thought reasoning or other methods, leveraging the similarity of user messages to their emails can give insight into their learning and indicate a way to improve training by adjusting the teaching approach in response to these types of messages. In the dataset we present, we noted a correlation between teacher and student message similarity with respect to several metics, which indicates that LLM teachers are often similarly narrow-focused as students. The degree of this specificity could be adjusted in response to student message similarity to emails, and avoid merely mirroring the specificity that user messages exhibit.

In addition to the significant positive correlations we report, there are also interesting negative correlations that differ from expectations given the correlation of other demographics and educational metrics. Specifically, we found that students who frequently send messages that are more closely related to the emails being observed actually had worse overall performance and training improvement. This can be explained by several different causes, such as less knowledgeable students more often choosing to ask questions that make reference to specific aspects of the emails they are observing, rather than the topic they are learning more broadly. This type of effect may allow for a chain of thought reasoning LLM model to identify when students are sending messages of this type, and adjust the method of providing educational feedback based on this insight.

By implementing these recommendations, anti-phishing and other types of online training platforms that use LLMs can potentially produce more responsive educational tools rather than one-size-fits-all chatbots that could disproportionally negatively impact the educational quality of important subpopulations. However, there are limitations to this work that raise important areas for future research. As mentioned, we performed only regression and mediation analysis on the demographics and learning outcomes of the dataset we had available, and our introduced embeddings of conversations. While this allowed us to make useful recommendations for future LLM teaching models, it is a limited view of the ways that LLM models can be improved. One useful area of future research that could leverage this same dataset or collect new data would be to compare the prompting of the LLMs and how they output educational feedback. LLM prompting was not a major investigation of this research as we chose to create embeddings of messages themselves, but a similar approach using LLM prompts could also be used to draw conclusions for important targets of LLM teacher optimization.

Beyond the work we present here, there are many additional contexts that LLM teaching feedback improvement can be applied to. Educational settings are one high-risk application of LLMs, which requires significant research into improving response quality and ensuring a lack of bias. Part of the reason for this is that in many situations humans will be interacting directly with the LLM without a dedicated human teacher. Alternative settings may have lower risks associated with them, such as in teaming settings where humans are using LLMs in cybersecurity contexts such as paired programming or as a tool for network analysis, threat detection, and a variety of other applications. Further research into how the results here can be applied to these settings can add to our understanding of how LLMs interact with humans.

Ethics Statement

The use of large language models (LLMs) in education carries significant ethical challenges. LLM outputs can be impacted by existing societal biases, such as those related to race, gender, or age. These biases have the potential to cause unequal learning experiences or reinforcing harmful stereotypes. The dataset presented in this work, and the recommendations we give to future LLM teaching models, are intended to mitigate the issues associated with unequal learning outcomes through our analysis of the learning of specific subpopulations and how it can be improved.

Training the LLM models used to converse with human participants, as well as the embedding models used to create the dataset we present demands immense computational resources, contributing to carbon emissions and other environmental impacts. Moreover, many widely used LLMs are built on datasets that include text gathered without the creators' knowledge or consent, raising serious questions about intellectual property rights, privacy, and the equitable sharing of benefits from such data. We attempted to mitigate these concerns through our analysis of different open and closed source embedding models in their effectiveness to relate embeddings to student learning outcomes. We additionally compared embedding models of different sizes to evaluate how smaller less computational intensive models fair in our applications.

Author Contributions

Tailia Malloy: conceptualization, data analysis, manuscript writing, manuscript editing, Laura Bernardy: conceptualization, manuscript writing, manuscript editing, Omar El Bachyr: conceptualization, manuscript editing, Fred Philippy: conceptualization, manuscript editing, Jordan Samhi: conceptualization, manuscript editing, Jacques Klein: conceptualization, manuscript editing, funding acquisition, Tegawendé F. Bissyandé: conceptualization, manuscript editing, funding acquisition.

Data Availability Statement

All data including models, datasets, and code used to plot figures, and generate datasets is included in the linked OSF and github repositories.

Conflicts of Interest/Disclaimer

This work was performed in collaboration with the private company Zortify S.A 19, rue du Labratoire L-1911 Luxembourg

Funding Acknowledgments

This research was funded by the University of Luxembourg Interdisciplinary Center for Security, Reliability and Trust research project titled "AI4HR Explainable psychometric AI for HR decision support (AI4HR)" in collaboration with Zortify S.A. 19, rue du Labratoire L-1911 Luxembourg.

Ethics Review Board Statement

The dataset presented in this work is based off of an existing dataset. Additional information on the preregistration and IRB is available on the existing OSF repository[11]. That experiment was preregistered on OSF, the experiment was approved by the Internal Review Board of Carnegie Mellon University and all participants were provided with information on their participation and the benefits and risks to them.

[11] https://www.osf.io/wbg3r/

Bibliography

1. Salemi, A.; Mysore, S.; Bendersky, M.; Zamani, H. LaMP: When Large Language Models Meet Personalization. *Arxiv* **2024**.
2. Woźniak, S.; Koptyra, B.; Janz, A.; Kazienko, P.; Kocoń, J. Personalized Large Language Models. *Arxiv* **2024**.
3. Zhang, Z.; Rossi, R.A.; Kveton, B.; Shao, Y.; Yang, D.; Zamani, H.; Dernoncourt, F.; Barrow, J.; Yu, T.; Kim, S.; et al. Personalization of Large Language Models: A Survey. *Arxiv* **2025**.
4. Malloy, T.; Ferriera, M.; Fang, F.; Gonzalez, C. Improving Online Anti-Phishing Training Using Cognitive Large Language Models. *In Press for Computers and Human Behavior* **2025**.
5. Pedersen, K.T.; Pepke, L.; Stærmose, T.; Papaioannou, M.; Choudhary, G.; Dragoni, N. Deepfake-Driven Social Engineering: Threats, Detection Techniques, and Defensive Strategies in Corporate Environments. *Journal of Cybersecurity and Privacy* **2025**, *5*, 18.
6. Schmitt, M.; Flechais, I. Digital deception: Generative artificial intelligence in social engineering and phishing. *Artificial Intelligence Review* **2024**, *57*, 324.
7. Jabir, R.; Le, J.; Nguyen, C. Phishing Attacks in the Age of Generative Artificial Intelligence: A Systematic Review of Human Factors. *AI* **2025**, *6*, 174.
8. Ayodele, T.O. Impact of AI-Generated Phishing Attacks: A New Cybersecurity Threat. In Proceedings of the Intelligent Computing-Proceedings of the Computing Conference. Springer, 2025, pp. 301–320.
9. Mahal, A.; Singh, K.; Singh, K. Influence of Generative AI on Cyber Security. *International journal of all research education and scientific methods* **2025**, pp. 1922–1928.
10. Proofpoint. 2024 State of the Phish: Risky actions, real-world threats and user resilience in an age of human-centric cybersecurity.
11. Hartzler, B.; Hinde, J.; Lang, S.; Correia, N.; Yermash, J.; Yap, K.; Murphy, C.M.; Ruwala, R.; Rash, C.J.; Becker, S.J.; et al. Virtual training is more cost-effective than in-person training for preparing staff to implement contingency management. *Journal of Technology in Behavioral Science* **2023**, *8*, 255–264.
12. o'Doherty, D.; Dromey, M.; Lougheed, J.; Hannigan, A.; Last, J.; McGrath, D. Barriers and solutions to online learning in medical education–an integrative review. *BMC medical education* **2018**, *18*, 130.
13. Luo, Y.; Yang, Y. Large language model and domain-specific model collaboration for smart education. *Frontiers of Information Technology & Electronic Engineering* **2024**, *25*, 333–341.
14. Neumann, A.T.; Yin, Y.; Sowe, S.; Decker, S.; Jarke, M. An llm-driven chatbot in higher education for databases and information systems. *IEEE Transactions on Education* **2024**.
15. Hang, C.N.; Tan, C.W.; Yu, P.D. MCQGen: A large language model-driven MCQ generator for personalized learning. *IEEE Access* **2024**, *12*, 102261–102273.
16. Chu, Z.; Wang, S.; Xie, J.; Zhu, T.; Yan, Y.; Ye, J.; Zhong, A.; Hu, X.; Liang, J.; Yu, P.S.; et al. Llm agents for education: Advances and applications. *arXiv preprint arXiv:2503.11733* **2025**.
17. Plaat, A.; van Duijn, M.; van Stein, N.; Preuss, M.; van der Putten, P.; Batenburg, K.J. Agentic large language models, a survey. *arXiv preprint arXiv:2503.23037* **2025**.
18. Ranjan, R.; Gupta, S.; Singh, S.N. A comprehensive survey of bias in llms: Current landscape and future directions. *arXiv preprint arXiv:2409.16430* **2024**.
19. Dong, X.; Wang, Y.; Yu, P.S.; Caverlee, J. Disclosure and mitigation of gender bias in llms. *arXiv preprint arXiv:2402.11190* **2024**.
20. Duan, Y.; Tang, F.; Wu, K.; Guo, Z.; Huang, S.; Mei, Y.; Wang, Y.; Yang, Z.; Gong, S. The large language model (LLM) bias evaluation (age bias). *DIKWP Research Group International Standard Evaluation. DOI* **2024**, *10*.
21. Weissburg, I.; Anand, S.; Levy, S.; Jeong, H. Llms are biased teachers: Evaluating llm bias in personalized education. *arXiv preprint arXiv:2410.14012* **2024**.
22. Liu, J.; Qiu, Z.; Li, Z.; Dai, Q.; Zhu, J.; Hu, M.; Yang, M.; King, I. A Survey of Personalized Large Language Models: Progress and Future Directions. *Arxiv* **2025**.

23. Islam, S.B.; Rahman, M.A.; Hossain, K.S.M.T.; Hoque, E.; Joty, S.; Parvez, M.R. Open-RAG: Enhanced Retrieval-Augmented Reasoning with Open-Source Large Language Models. *Arxiv* **2024**.
24. Gao, L.; Lu, J.; Shao, Z.; Lin, Z.; Yue, S.; Leong, C.; Sun, Y.; Zauner, R.J.; Wei, Z.; Chen, S. Fine-tuned large language model for visualization system: A study on self-regulated learning in education. *IEEE Transactions on Visualization and Computer Graphics* **2024**, *31*, 514–524.
25. Salemi, A.; Zamani, H. Comparing retrieval-augmentation and parameter-efficient fine-tuning for privacy-preserving personalization of large language models. In Proceedings of the Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR), 2025, pp. 286–296.
26. Jin, C.; Zhang, Z.; Jiang, X.; Liu, F.; Liu, X.; Liu, X.; Jin, X. RAGCache: Efficient Knowledge Caching for Retrieval-Augmented Generation. *Arxiv* **2024**.
27. Sun, C.; Yang, K.; Reddy, R.G.; Fung, Y.R.; Chan, H.P.; Small, K.; Zhai, C.; Ji, H. Persona-DB: Efficient Large Language Model Personalization for Response Prediction with Collaborative Data Refinement. *Arxiv* **2025**.
28. Yu, Z.; Liu, S.; Denny, P.; Bergen, A.; Liut, M. Integrating small language models with retrieval-augmented generation in computing education: Key takeaways, setup, and practical insights. In Proceedings of the Proceedings of the 56th ACM Technical Symposium on Computer Science Education V. 1, 2025, pp. 1302–1308.
29. Sanyal, D.; Maiti, A.; Maharana, U.; Kumar, D.; Mali, A.; Giles, C.L.; Mandal, M. Investigating Pedagogical Teacher and Student LLM Agents: Genetic Adaptation Meets Retrieval Augmented Generation Across Learning Style. *arXiv preprint arXiv:2505.19173* **2025**.
30. Afane, K.; Wei, W.; Mao, Y.; Farooq, J.; Chen, J. Next-Generation Phishing: How LLM Agents Empower Cyber Attackers, 2024, [arXiv:cs.CR/2411.13874].
31. Bethany, M.; Galiopoulos, A.; Bethany, E.; Karkevandi, M.B.; Vishwamitra, N.; Najafirad, P. Large language model lateral spear phishing: A comparative study in large-scale organizational settings. *arXiv preprint arXiv:2401.09727* **2024**.
32. Google Cloud. Google Cloud Cybersecurity Forecast 2024. https://services.google.com/fh/files/misc/google-cloud-cybersecurity-forecast-2024.pdf, 2024. Accessed: 2025-08-08.
33. Sharma, M.; Singh, K.; Aggarwal, P.; Dutt, V. How well does GPT phish people? An investigation involving cognitive biases and feedback. In Proceedings of the 2023 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW). IEEE, 2023, pp. 451–457.
34. Hasanov, I.; Virtanen, S.; Hakkala, A.; Isoaho, J. Application of large language models in cybersecurity: A systematic literature review. *IEEE Access* **2024**.
35. Hazell, J. Spear Phishing With Large Language Models, 2023, [arXiv:cs.CY/2305.06972].
36. Myung, J.; Lee, N.; Zhou, Y.; Jin, J.; Putri, R.; Antypas, D.; Borkakoty, H.; Kim, E.; Perez-Almendros, C.; Ayele, A.A.; et al. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. *Advances in Neural Information Processing Systems* **2024**, *37*, 78104–78146.
37. Jampen, D.; Gür, G.; Sutter, T.; Tellenbach, B. Don't click: towards an effective anti-phishing training. A comparative literature review. *Human-centric Computing and Information Sciences* **2020**, *10*, 33.
38. Huang, L.; Jia, S.; Balcetis, E.; Zhu, Q. Advert: an adaptive and data-driven attention enhancement mechanism for phishing prevention. *IEEE Transactions on Information Forensics and Security* **2022**, *17*, 2585–2597.
39. Singh, K.; Aggarwal, P.; Rajivan, P.; Gonzalez, C. Training to detect phishing emails: Effects of the frequency of experienced phishing emails. In Proceedings of the Proceedings of the human factors and ergonomics society annual meeting. SAGE Publications Sage CA: Los Angeles, CA, 2019, Vol. 63, pp. 453–457.
40. Brunken, L.; Buckmann, A.; Hielscher, J.; Sasse, M.A. {"To} Do This Properly, You Need More {Resources"}: The Hidden Costs of Introducing Simulated Phishing Campaigns. In Proceedings of the 32nd USENIX Security Symposium (USENIX Security 23), 2023, pp. 4105–4122.
41. Salahdine, F.; Kaabouch, N. Social engineering attacks: A survey. *Future internet* **2019**, *11*, 89.

42. Abroshan, H.; Devos, J.; Poels, G.; Laermans, E. Phishing happens beyond technology: The effects of human behaviors and demographics on each step of a phishing process. *IEEE Access* **2021**, *9*, 44928–44949.
43. Singh, K.; Aggarwal, P.; Rajivan, P.; Gonzalez, C. Cognitive elements of learning and discriminability in anti-phishing training. *Computers & Security* **2023**, *127*, 103105.
44. Malloy, T.; Gonzalez, C. Applying Generative Artificial Intelligence to Cognitive Models of Decision Making. *Frontiers in Psychology* **2024**.
45. OpenAI. OpenAI API Documentation. https://platform.openai.com/docs/, 2025. Accessed: May 18, 2025.
46. Gonzalez, C.; Lerch, J.F.; Lebiere, C. Instance-based learning in dynamic decision making. *Cognitive Science* **2003**, *27*, 591–635.
47. Gonzalez, C.; Dutt, V. Instance-based learning: integrating sampling and repeated decisions from experience. *Psychological review* **2011**, *118*, 523.
48. Gonzalez, C. Building Human-Like Artificial Agents: A General Cognitive Algorithm for Emulating Human Decision-Making in Dynamic Environments. *Perspectives on Psychological Science* **2023**, p. 17456916231196766.
49. Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388* **2025**.
50. Xiao, S.; Liu, Z.; Zhang, P.; Muennighoff, N. C-Pack: Packaged Resources To Advance General Chinese Embedding, 2023, [arXiv:cs.CL/2309.07597].
51. Team, G.; Kamath, A.; Ferret, J.; Pathak, S.; Vieillard, N.; Merhej, R.; Perrin, S.; Matejovicova, T.; Ramé, A.; Rivière, M.; et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786* **2025**.
52. Awasthy, P.; Trivedi, A.; Li, Y.; Doshi, M.; Bhat, R.; Kumar, V.; Yang, Y.; Iyer, B.; Daniels, A.; Murthy, R.; et al. Granite Embedding R2 Models. *arXiv preprint arXiv:2508.21085* **2025**.
53. Reimers, N.; Gurevych, I. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084* **2019**.
54. Wang, Z.; Ning, X.; Blaschko, M. Jaccard metric losses: Optimizing the jaccard index with soft labels. *Advances in Neural Information Processing Systems* **2023**, *36*, 75259–75285.
55. Lin, C.Y.; Och, F. Looking for a few good metrics: ROUGE and its evaluation. In Proceedings of the Ntcir workshop, 2004, pp. 1–8.
56. Huang, A.; et al. Similarity measures for text document clustering. In Proceedings of the Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand, 2008, Vol. 4, pp. 9–56.
57. Li, B.; Zhou, H.; He, J.; Wang, M.; Yang, Y.; Li, L. On the sentence embeddings from pre-trained language models. *arXiv preprint arXiv:2011.05864* **2020**.
58. Fiedler, K.; Schott, M.; Meiser, T. What mediation analysis can (not) do. *Journal of Experimental Social Psychology* **2011**, *47*, 1231–1236.
59. Baron, R.M.; Kenny, D.A. The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of personality and social psychology* **1986**, *51*, 1173.
60. Bodker, S.; Andersen, P.B. Complex mediation. *Human-computer interaction* **2005**, *20*, 353–402.
61. Vig, J.; Gehrmann, S.; Belinkov, Y.; Qian, S.; Nevo, D.; Singer, Y.; Shieber, S. Investigating gender bias in language models using causal mediation analysis. *Advances in neural information processing systems* **2020**, *33*, 12388–12401.
62. Vallat, R. Pingouin: statistics in Python. *Journal of Open Source Software* **2018**, *3*, 1026. https://doi.org/10.21105/joss.01026.
63. Alwanain, M.I. Phishing awareness and elderly users in social media. *International Journal of Computer Science and Network Security* **2020**, *20*, 114–119.

Appendix

In this appendix, ME-CS refers to the email cosine similarity metric used in our regression analyses.

ANOVA analyses tables

Table 2. ANOVA results relating ME-CS to outcomes

Source	Outcome	df_1	df_2	F	p	η_p^2
Student	Correct	22	464	0.841	0.674	0.038
Teacher	Correct	25	1720	1.648	0.0231	0.023
Student	Confidence	22	464	1.539	0.0569	0.068
Teacher	Confidence	25	1720	1.652	0.0225	0.023
Student	ReactionTime	22	464	1.155	0.284	0.052
Teacher	ReactionTime	25	1720	0.882	0.632	0.013
Student	User Initial Performance	22	464	0.692	0.849	0.032
Teacher	User Initial Performance	25	1720	0.863	0.659	0.012
Student	User Improvement	22	464	1.557	0.0521	0.069
Teacher	User Improvement	25	1720	1.014	0.444	0.015
Student	User Final Performance	22	464	1.705	0.0247	0.075
Teacher	User Final Performance	25	1720	1.189	0.237	0.017
Student	Pre-Experiment Quiz Score	22	464	1.195	0.247	0.054
Teacher	Pre-Experiment Quiz Score	25	1720	1.261	0.174	0.018
Student	AI Gen Percept	22	464	1.348	0.135	0.060
Teacher	AI Gen Percept	25	1720	0.702	0.86	0.010
Student	Response Mssg Sim	22	464	5.624	4.86e−14	0.211
Teacher	Response Mssg Sim	25	1720	1.377	0.102	0.020
Student	Age	22	464	1.395	0.11	0.062
Teacher	Age	25	1720	1.122	0.307	0.016
Student	Gender Number	22	464	1.110	0.331	0.050
Teacher	Gender Number	25	1720	0.880	0.635	0.013
Student	Education Years	22	464	0.991	0.474	0.045
Teacher	Education Years	25	1720	0.984	0.486	0.014
Student	Phishing Experience	22	464	0.923	0.565	0.042
Teacher	Phishing Experience	25	1720	0.912	0.589	0.013
Student	Chatbot Experience	22	464	1.332	0.144	0.059
Teacher	Chatbot Experience	25	1720	1.016	0.442	0.015
Student	Cognitive Model Activity	22	464	1.725	0.0222	0.076
Teacher	Cognitive Model Activity	25	1720	1.159	0.267	0.017

Mediation Analyses

Table 3. Mediation analysis Student and Teacher Messages on User Improvement by Age

Path	Coef.	SE	p	CI 2.5%	CI 97.5%	Sig
Age						
∼ Age	-0.0609	0.0211	0.00396	-0.102	-0.0195	Yes
User Improvement						
∼ Age	0.00136	0.0212	0.949	-0.0402	0.0429	No

Path	Coef.	SE	p	CI 2.5%	CI 97.5%	Sig
Total	-0.0264	0.0212	0.212	-0.0679	0.0151	No
Direct	-0.0264	0.0212	0.213	-0.068	0.0151	No
Indirect	1.507e−05	0.00116	0.98	-0.00254	0.00219	No

Table 4. Mediation analysis Student and Teacher Messages on User Improvement by Education Years

Path	Coef.	SE	p	CI 2.5%	CI 97.5%	Sig
Education Years						
∼ Education Years	-0.0326	0.0212	0.124	-0.0741	0.00893	No
User Improvement						
∼ Education Years	-0.214	0.0207	1.550e−24	-0.255	-0.173	Yes
Total	-0.0264	0.0212	0.212	-0.0679	0.0151	No
Direct	-0.0334	0.0207	0.106	-0.074	0.00713	No
Indirect	0.007	0.00466	0.1	-0.00167	0.017	No

Table 5. Mediation analysis Age Messages on Correct Categorization by Student and Teacher

Path	Coef.	SE	p	CI 2.5%	CI 97.5%	Sig
ME-CS						
∼ Age	-0.0609	0.0211	0.00396	-0.102	-0.0195	Yes
User Improvement						
∼ ME-CS	0.093	0.0211	1.082e−05	0.0516	0.134	Yes
Total	0.046	0.0211	0.0299	0.00448	0.0874	Yes
Direct	0.0518	0.0211	0.0141	0.0104	0.0932	Yes
Indirect	-0.00586	0.00238	0.012	-0.0109	-0.0019	Yes

Table 6. Mediation analysis Education Years Messages on Correct Categorization by Student and Teacher

Path	Coef.	SE	p	CI 2.5%	CI 97.5%	Sig
ME-CS						
∼ Education Years	-0.0326	0.0212	0.124	-0.0741	0.00893	No
User Improvement						
∼ ME-CS	0.093	0.0211	1.082e−05	0.0516	0.134	Yes
Total	-0.151	0.0209	6.906e−13	-0.192	-0.11	Yes
Direct	-0.148	0.0209	1.561e−12	-0.189	-0.107	Yes
Indirect	-0.00287	0.00208	0.1	-0.0085	0.000427	No

Table 7. Mediation analysis Phishing Experience Messages on Correct Categorization by Student and Teacher

Path	Coef.	SE	p	CI 2.5%	CI 97.5%	Sig
ME-CS						
∼ Phishing Experience	0.0242	0.0212	0.254	-0.0173	0.0657	No

Path	Coef.	SE	p	CI 2.5%	CI 97.5%	Sig
User Improvement						
∼ ME-CS	0.093	0.0211	1.082e−05	0.0516	0.134	Yes
Total	0.172	0.0209	2.797e−16	0.131	0.213	Yes
Direct	0.17	0.0208	5.082e−16	0.129	0.211	Yes
Indirect	0.00215	0.00202	0.26	-0.00159	0.0064	No

Table 8. Mediation analysis Chatbot Experience Messages on Correct Categorization by Student and Teacher

Path	Coef.	SE	p	CI 2.5%	CI 97.5%	Sig
ME-CS						
∼ Chatbot Experience	0.000442	0.0212	0.983	-0.0411	0.042	No
User Improvement						
∼ ME-CS	0.093	0.0211	1.082e−05	0.0516	0.134	Yes
Total	-0.0209	0.0212	0.323	-0.0624	0.0206	No
Direct	-0.021	0.0211	0.32	-0.0623	0.0204	No
Indirect	4.114e−05	0.00202	0.94	-0.00374	0.00476	No

Table 9. Mediation analysis AI Generation Perception Messages on Correct Categorization by Student and Teacher

Path	Coef.	SE	p	CI 2.5%	CI 97.5%	Sig
ME-CS						
∼ AI Generation Perception	0.0403	0.0212	0.0567	-0.00115	0.0818	No
User Improvement						
∼ ME-CS	0.093	0.0211	1.082e−05	0.0516	0.134	Yes
Total	-0.12	0.021	1.293e−08	-0.161	-0.0788	Yes
Direct	-0.124	0.0209	3.741e−09	-0.165	-0.0829	Yes
Indirect	0.00395	0.0022	0.044	0.000156	0.00834	Yes

Table 10. Mediation analysis Pre Experiment Quiz Score Messages on Correct Categorization by Student and Teacher

Path	Coef.	SE	p	CI 2.5%	CI 97.5%	Sig
ME-CS						
∼ Pre Experiment Quiz Score	-0.0079	0.0212	0.709	-0.0494	0.0336	No
User Improvement						
∼ ME-CS	0.093	0.0211	1.082e−05	0.0516	0.134	Yes
Total	0.0702	0.0211	9.064e−04	0.0288	0.112	Yes
Direct	0.0709	0.021	7.603e−04	0.0297	0.112	Yes
Indirect	-0.000738	0.00204	0.684	-0.0046	0.00331	No

Table 11. Mediation analysis Response Message Similarity Messages on Correct Categorization by Student and Teacher

Path	Coef.	SE	p	CI 2.5%	CI 97.5%	Sig
ME-CS						
∼ Response Message Similarity	0.149	0.0209	1.244e−12	0.108	0.191	Yes
User Improvement						
∼ ME-CS	0.093	0.0211	1.082e−05	0.0516	0.134	Yes
Total	0.235	0.0206	2.062e−29	0.195	0.275	Yes
Direct	0.226	0.0208	6.212e−27	0.185	0.267	Yes
Indirect	0.00884	0.00347	0.004	0.00285	0.0157	Yes

Table 12. Mediation analysis Age Messages on Correct Categorization by Teacher

Path	Coef.	SE	p	CI 2.5%	CI 97.5%	Sig
ME-CS						
∼ Age	-0.0347	0.0239	0.147	-0.0817	0.0122	No
User Improvement						
∼ ME-CS	0.109	0.0238	4.827e−06	0.0625	0.156	Yes
Total	0.0323	0.0239	0.177	-0.0146	0.0793	No
Direct	0.0362	0.0238	0.129	-0.0105	0.0829	No
Indirect	-0.00383	0.00258	0.1	-0.00978	0.00013	No

Table 13. Mediation analysis Education Years Messages on Correct Categorization by Teacher

Path	Coef.	SE	p	CI 2.5%	CI 97.5%	Sig
ME-CS						
∼ Education Years	-0.0563	0.0239	0.0186	-0.103	-0.00944	Yes
User Improvement						
∼ ME-CS	0.109	0.0238	4.827e−06	0.0625	0.156	Yes
Total	-0.109	0.0238	5.421e−06	-0.155	-0.0619	Yes
Direct	-0.103	0.0237	1.557e−05	-0.149	-0.0562	Yes
Indirect	-0.00582	0.00292	0.02	-0.0131	-0.00137	Yes

Table 14. Mediation analysis Phishing Experience Messages on Correct Categorization by Teacher

Path	Coef.	SE	p	CI 2.5%	CI 97.5%	Sig
ME-CS						
∼ Phishing Experience	-0.00269	0.0239	0.91	-0.0497	0.0443	No
User Improvement						
∼ ME-CS	0.109	0.0238	4.827e−06	0.0625	0.156	Yes
Total	0.0798	0.0239	8.405e−04	0.033	0.127	Yes
Direct	0.0801	0.0237	7.498e−04	0.0336	0.127	Yes
Indirect	-0.000295	0.00264	0.848	-0.00468	0.00516	No

Table 15. Mediation analysis Chatbot Experience Messages on Correct Categorization by Teacher

Path	Coef.	SE	p	CI 2.5%	CI 97.5%	Sig
ME-CS						
∼ Chatbot Experience	0.0254	0.0239	0.288	-0.0215	0.0724	No
User Improvement						
∼ ME-CS	0.109	0.0238	4.827e−06	0.0625	0.156	Yes
Total	-0.0248	0.0239	0.301	-0.0717	0.0222	No
Direct	-0.0275	0.0238	0.247	-0.0742	0.0192	No
Indirect	0.00279	0.00288	0.316	-0.00189	0.00933	No

Table 16. Mediation analysis AI Generation Perception Messages on Correct Categorization by Teacher

Path	Coef.	SE	p	CI 2.5%	CI 97.5%	Sig
ME-CS						
∼ AI Generation Perception	0.00958	0.0239	0.689	-0.0374	0.0565	No
User Improvement						
∼ ME-CS	0.109	0.0238	4.827e−06	0.0625	0.156	Yes
Total	-0.082	0.0239	6.084e−04	-0.129	-0.0351	Yes
Direct	-0.083	0.0237	4.797e−04	-0.13	-0.0365	Yes
Indirect	0.00105	0.00283	0.756	-0.00371	0.00719	No

Table 17. Mediation analysis Pre Experiment Quiz Score Messages on Correct Categorization by Teacher

Path	Coef.	SE	p	CI 2.5%	CI 97.5%	Sig
ME-CS						
∼ Pre Experiment Quiz Score	0.00486	0.0239	0.839	-0.0421	0.0518	No
User Improvement						
∼ ME-CS	0.109	0.0238	4.827e−06	0.0625	0.156	Yes
Total	0.052	0.0239	0.0298	0.00511	0.0989	Yes
Direct	0.0515	0.0238	0.0305	0.00484	0.0981	Yes
Indirect	0.000529	0.00242	0.88	-0.0038	0.00657	No

Table 18. Mediation analysis Response Message Similarity Messages on Correct Categorization by Teacher

Path	Coef.	SE	p	CI 2.5%	CI 97.5%	Sig
ME-CS						
∼ Response Message Similarity	0.106	0.0238	9.378e−06	0.0591	0.153	Yes
User Improvement						
∼ ME-CS	0.109	0.0238	4.827e−06	0.0625	0.156	Yes
Total	0.217	0.0234	4.474e−20	0.171	0.263	Yes
Direct	0.208	0.0234	1.665e−18	0.162	0.254	Yes
Indirect	0.00922	0.00296	0.0e+00	0.00459	0.0165	Yes

Table 19. Mediation analysis Age Messages on Correct Categorization by Student

Path	Coef.	SE	p	CI 2.5%	CI 97.5%	Sig
ME-CS						
∼ Age	-0.131	0.045	0.00371	-0.22	-0.0428	Yes
User Improvement						
∼ ME-CS	0.0628	0.0453	0.166	-0.0262	0.152	No
Total	0.092	0.0452	0.0424	0.00316	0.181	Yes
Direct	0.102	0.0455	0.0255	0.0126	0.191	Yes
Indirect	-0.01	0.00724	0.104	-0.0292	0.000972	No

Table 20. Mediation analysis Education Years Messages on Correct Categorization by Student

Path	Coef.	SE	p	CI 2.5%	CI 97.5%	Sig
ME-CS						
∼ Education Years	0.0234	0.0454	0.607	-0.0658	0.113	No
User Improvement						
∼ ME-CS	0.0628	0.0453	0.166	-0.0262	0.152	No
Total	-0.247	0.044	3.563e−08	-0.333	-0.16	Yes
Direct	-0.248	0.044	2.820e−08	-0.334	-0.162	Yes
Indirect	0.0016	0.00366	0.58	-0.00317	0.0116	No

Table 21. Mediation analysis Phishing Experience Messages on Correct Categorization by Student

Path	Coef.	SE	p	CI 2.5%	CI 97.5%	Sig
ME-CS						
∼ Phishing Experience	0.0839	0.0452	0.0642	-0.00498	0.173	No
User Improvement						
∼ ME-CS	0.0628	0.0453	0.166	-0.0262	0.152	No
Total	0.353	0.0425	1.063e−15	0.269	0.436	Yes
Direct	0.35	0.0427	2.197e−15	0.266	0.434	Yes
Indirect	0.00281	0.00433	0.5	-0.00345	0.0142	No

Table 22. Mediation analysis Chatbot Experience Messages on Correct Categorization by Student

Path	Coef.	SE	p	CI 2.5%	CI 97.5%	Sig
ME-CS						
∼ Chatbot Experience	-0.0704	0.0453	0.121	-0.159	0.0186	No
User Improvement						
∼ ME-CS	0.0628	0.0453	0.166	-0.0262	0.152	No
Total	-0.000662	0.0454	0.988	-0.0899	0.0886	No
Direct	0.00378	0.0455	0.934	-0.0856	0.0931	No
Indirect	-0.00444	0.00473	0.264	-0.0177	0.00179	No

Table 23. Mediation analysis AI Generation Perception Messages on Correct Categorization by Student

Path	Coef.	SE	p	CI 2.5%	CI 97.5%	Sig
ME-CS						
∼ AI Generation Perception	0.114	0.0451	0.0121	0.025	0.202	Yes
User Improvement						
∼ ME-CS	0.0628	0.0453	0.166	-0.0262	0.152	No
Total	-0.192	0.0446	2.017e−05	-0.279	-0.104	Yes
Direct	-0.202	0.0447	8.245e−06	-0.29	-0.114	Yes
Indirect	0.00975	0.00632	0.06	0.00106	0.0258	No

Table 24. Mediation analysis Pre Experiment Quiz Score Messages on Correct Categorization by Student

Path	Coef.	SE	p	CI 2.5%	CI 97.5%	Sig
ME-CS						
∼ Pre Experiment Quiz Score	-0.0423	0.0454	0.351	-0.131	0.0468	No
User Improvement						
∼ ME-CS	0.0628	0.0453	0.166	-0.0262	0.152	No
Total	0.112	0.0451	0.0136	0.0231	0.2	Yes
Direct	0.115	0.0451	0.0114	0.026	0.203	Yes
Indirect	-0.00287	0.00459	0.48	-0.0177	0.0027	No

Table 25. Mediation analysis Response Message Similarity Messages on Correct Categorization by Student

Path	Coef.	SE	p	CI 2.5%	CI 97.5%	Sig
ME-CS						
∼ Response Message Similarity	0.336	0.0428	2.460e−14	0.252	0.42	Yes
User Improvement						
∼ ME-CS	0.0628	0.0453	0.166	-0.0262	0.152	No
Total	0.201	0.0445	7.485e−06	0.114	0.289	Yes
Direct	0.203	0.0473	2.069e−05	0.11	0.296	Yes
Indirect	-0.00185	0.0162	0.912	-0.0355	0.0283	No

Pre-experiment Instructions

Instructions. In this experiment you will determine whether example emails are genuine or phishing. When reviewing potential phishing emails, pay attention to the following features. After this screen, there will be a quiz on this information.

- **Real sender does not match the claimed sender:** Phishing emails often pretend to be from reputable companies, but you can usually spot a fake by checking the address that sent the message. If the `From` address is a series of numbers, an odd mix of characters, or not the official domain of the company it claims to be from, it's likely a phishing attempt.
- **Email requests credentials:** Legitimate companies will *never* ask for sensitive information via email. If the email requests your username, password, credit card information, or other sensitive data, it's a phishing attempt.

- **Suspicious subject line:** Phishing emails often use alarmist, threatening, or enticing subject lines to grab your attention. If the subject is odd, generic, or doesn't match the content, it could be a phishing email.
- **Urgent tone:** Phishing scams create a sense of urgency to panic you into acting without thinking. If an email asks for immediate action (e.g., "Your account will be suspended unless you update your information"), it's likely a scam.
- **Too-good-to-be-true offers:** Emails that promise rewards, discounts, or prizes in exchange for personal information are likely phishing.
- **Link does not match the text:** A common tactic is disguising a dangerous link with innocent-looking text. Hover your cursor over links before clicking. If the URL doesn't match the link text, or looks suspicious in any way, do not click. For instance, if the link text reads "bank.com" but hovering shows "hackingsite.com", it's a phishing attempt.

Pre-experiment Quiz

1. What type of language do phishing emails often use to create a sense of panic?
    - Urgent language
    - Friendly language
    - Rude language
    - Mean language
2. What might a phishing email request of you that would compromise your identity?
    - Personal information like your favorite color
    - Sensitive information like credit card numbers
    - Sensitive information like your celebrity crush
    - Irrelevant information like your dog's name
3. What types of actions might phishing emails request from you that could lead to malware being installed on your computer?
    - Clicking links only
    - Downloading attachments only
    - Replying with your computer's information only
    - All of the above
4. How might a phishing email try to ensure that you are susceptible to a phishing attempt?
    - Being overly friendly
    - Calling you a generic title
    - Using poor grammar
    - Saying you won the lottery
5. How might a phishing email attempt to convince you that it was sent from a legitimate source?
    - Using an email from a website that you have never heard of
    - Sending the email from a website with a famous company name
    - Adding a link to a real website in the text of the email
    - Using another website name that is different from the one sending the email
6. How might a phishing email convince you to click on a fake link?
    - Adding a lot of random numbers and letters into the link
    - Changing the text of the link (can be checked by hovering over it)
    - Changing the color of the link to make it look like you've clicked it before
    - Keeping the link short so it looks legitimate

Experiment Questions

1. Is this a phishing email?
   - Yes
   - No
2. On a scale from 1–5, with 5 being totally confident, how confident are you in your answer to Question 1?
   - 1
   - 2
   - 3
   - 4
   - 5
3. What action would you take after receiving this email?
   - Respond
   - Click link
   - Check sender
   - Check link
   - Delete email
   - Report email

Post-experiment Questionnaire

1. Of the phishing emails you've encountered, what percentage do you think were generated by artificial intelligence models?
   - 100% of the phishing emails I read were written by an Artificial Intelligence model.
   - 75% of the phishing emails I read were written by an Artificial Intelligence model.
   - 50% of the phishing emails I read were written by an Artificial Intelligence model.
   - 25% of the phishing emails I read were written by an Artificial Intelligence model.
2. Of the ham (i.e., non-phishing) emails you've encountered, what percentage do you think were generated by artificial intelligence models?
   - 100% of the ham emails I read were written by an Artificial Intelligence model.
   - 75% of the ham emails I read were written by an Artificial Intelligence model.
   - 50% of the ham emails I read were written by an Artificial Intelligence model.
   - 25% of the ham emails I read were written by an Artificial Intelligence model.
3. Of the phishing emails you've encountered, what percentage do you think were *styled* (i.e., appearance and format) by artificial intelligence models?
   - 100% of the phishing emails I read were styled by an Artificial Intelligence model.
   - 75% of the phishing emails I read were styled by an Artificial Intelligence model.
   - 50% of the phishing emails I read were styled by an Artificial Intelligence model.
   - 25% of the phishing emails I read were styled by an Artificial Intelligence model.
4. Of the ham (i.e., non-phishing) emails you've encountered, what percentage do you think were *styled* (i.e., appearance and format) by artificial intelligence models?
   - 100% of the ham emails I read were styled by an Artificial Intelligence model.
   - 75% of the ham emails I read were styled by an Artificial Intelligence model.
   - 50% of the ham emails I read were styled by an Artificial Intelligence model.
   - 25% of the ham emails I read were styled by an Artificial Intelligence model.
5. What criteria did you use to identify whether an email was a phishing attempt? *Open response.*