Human Factors in Cybersecurity, Vol. 208, 2026, 42–53
https://doi.org/10.54941/ahfe1007411

Privileged Learning for Instance Representation in Cognitive Models of Phishing Decisions

Elaheh Mehrabi and Prashanth Rajivan

Industrial and Systems Engineering, Seattle, WA 98195, USA

ABSTRACT

Risk arising from human behavior, such as employees falling victim to phishing, continues to undermine organizational security posture. Prior work has attributed phishing susceptibility to attentional failures in detecting suspicious cues, motivating training approaches focused on detecting such cues. However, growing evidence suggests that susceptibility to phishing is better explained through activation and retrieval process of relevant experiences from memory. Models capable of estimating awareness gaps and predicting how individuals respond to or report phishing emails are therefore critical for delivering personalized training and testing interventions. A key challenge in building such cognitive models is finding effective ways to represent the contextual cues that shape how individuals perceive, store, and recall phishing-related content. This paper applies a privileged learning strategy to construct richer instance representations within cognitive models of phishing judgment. Combining instance-based learning (IBL) with neural network-based text similarity, we infer how recipients interpret email content and underlying intent. Results indicate that this privileged learning pipeline substantially enhances the predictive ability of cognitive models of phishing, opening new methods for developing individualized anti-phishing interventions.

Keywords: Phishing detection, Cognitive modeling, Privileged learning, Instance-based learning, Neural networks

INTRODUCTION

Phishing remains one of the most prevalent social engineering threats, exploiting human judgment to extract confidential data or trick recipients into taking harmful actions. Their growing prevalence mirrors the escalating scale and severity of cyberattacks (Gonzalez et al., 2014). Consequently, reducing risks arising from human (employee) behavior through security awareness training remains critical to the effectiveness of organizational technical defenses.

Early work attributed phishing susceptibility to human inattention, such as failing to notice suspicious URLs or sender addresses (Dhamija et al., 2006). As a result, many security awareness programs emphasize training employees to spot such cues. However, recent evidence shows that prevailing training and testing approaches yield limited and short-lived behavioral improvements (Ho et al., 2025). We argue that inattention is a symptom

Received January 30, 2026; Revised April 7, 2026; Accepted April 22, 2026; Available online July 20, 2026

rather than the root cause, arising when relevant security knowledge received through training is not reliably activated and retrieved while responding to phishing emails.

A growing body of work highlights that successful phishing detection depends on memory-driven processes: activating stored security knowledge and recalling relevant past encounters so that suspicious cues are recalled when needed (Cranford et al., 2019; Xu et al., 2022; Malloy and Gonzalez, 2024). Memory retrieval plays a central role because familiarity with particular cues, topics, or entities heavily influences how people judge whether an email is legitimate. Encountering references to known conversations or recognizable contacts can encourage trust by triggering easy recall of prior interactions, while well-known fraud patterns (e.g., "Nigerian scams") can automatically raise suspicion by activating stored knowledge about deceptive techniques.

Computational models grounded in ACT-R and Instance-Based Learning Theory (IBLT) offer a principled way to analyze how human memory processes influences email-related decisions (Cranford et al., 2019; Xu et al., 2022). A recurring difficulty, however, is instance engineering: determining what contextual features to include so that the model's internal representations mirror what individuals actually encode and later retrieve when judging emails (Xu and Rajivan, 2024). Earlier efforts employed transformer-based language models for this purpose, but treating an entire email as a single vector may not reflect human encoding of email content and can introduce error and bias (Xu et al., 2022; Devlin et al., 2018; Sannigrahi and Genabith, 2023). Studying how emails (including phishing emails) are stored and recalled is therefore essential for determining awareness gaps and developing personalized anti-phishing training and protection tools (Azuma et al., 2006; Newell and Simon, 1972).

We investigate Learning Using Privileged Information (LUPI) which is a training paradigm that allows models to exploit privileged data present only during the learning phase, not at deployment (Vapnik and Vashist, 2009). In our setting, the privileged data are survey responses gathered during laboratory experiments that record how participants interpreted key aspects of each email such as email intents (Vapnik and Vashist, 2009; Momeni et al., 2018). Although such self-reported data would not be available in operational environments, we hypothesize that using LUPI and incorporating privileged data during training for instance representations can be better predict how individuals will interpret new emails using cognitive models.

BACKGROUND

Beyond inattention to cues (Dhamija et al., 2006; Wu et al., 2006) and limited training effects (Kumaraguru et al., 2007; Kumaraguru et al., 2010), research links phishing vulnerability to social influence and memory-related inefficiencies that shape how people encode and later interpret threat cues (Cranford et al., 2019; Vishwanath et al., 2018; Ferreira and Teles, 2019; Sawyer and Hancock, 2018).

Attention and memory are tightly deeply coupled with each other. Becoming suspicious of an email often depends on an automatic process in which cues in phishing emails must trigger retrieval of relevant past experiences and knowledge. When retrieval is effective, people recognize patterns that suggest suspicion and adjust their behavior; when retrieval fails to surface relevant memories, cues may be processed as routine and the message may be treated as benign, increasing the chance of susceptibility (Anderson et al., 2004; Cranford et al., 2019).

According to IBLT, decisions are made from experience by recalling similar past experiences, the choices made then, and the payoffs received, to guide current choices (Gonzalez et al., 2003). Which memories surface depends on how recently and how often they were activated, as well as on the perceived resemblance between the present context and previously stored episodes (Anderson et al., 2004).

Earlier IBL-based phishing models encoded emails through email protocol-level attributes such as sender fields, subject lines, and similar metadata (Cranford et al., 2019; Cranford et al., 2021). Subsequent efforts embedded entire email bodies via language models, yet these dense, high-dimensional vectors proved challenging for IBL agents and likely diverge from the sparse, meaning-oriented traces that people actually retain after reading an email (Xu et al., 2022; Shonman et al., 2022). Evidence indicates that low-dimensional, survey-collected representations better capture the cues people genuinely recall, yielding more faithful cognitive models of phishing decisions (Xu et al., 2022). Nevertheless, obtaining such survey data is expensive and typically infeasible outside controlled experiments, leaving a gap between laboratory findings and real-world deployment.

Bridging this gap requires techniques that can leverage privilege data that captures how people encode phishing emails to memory which are more likely to be available during development yet operate without them at test time. LUPI (Vapnik and Vashist, 2009) is one technique that allows that were a model is first trained with auxiliary human annotations or survey cues available during training, and the resulting knowledge is distilled into predictions that rely solely on features accessible at deployment. Applied to phishing, LUPI offers a principled route for incorporating human-derived perceptual labels and annotations into the training process, improving both representation quality and generalization, while ensuring the final system needs nothing beyond the email text and metadata available in production settings.

METHOD

We describe a framework that couples IBL-based cognitive agents with sentence-transformer embeddings and feedforward neural networks to predict how individuals respond to both benign and phishing emails. We specifically evaluate LUPI (Vapnik and Vashist, 2009) as a means of producing lower-dimensional, human-relevant representations of email text for IBL models (Morrison and Gonzalez, 2024), comparing it against earlier approaches that relied on full email embeddings (Xu et al., 2022).

Dataset

Two datasets served distinct roles: Dataset 1 for model training and validation, Dataset 2 for out-of-sample evaluation. Dataset 1 originates from a controlled laboratory study on spear-phishing vulnerability (Xu et al., 2021). In that experiment, four-person groups operated within a simulated email environment: three members assumed employee roles guided by fictional backstories, while the fourth crafted phishing messages. Employees triaged a mixed inbox of legitimate, promotional, mass-phishing, and targeted spear-phishing emails. The dataset comprises 396 unique emails; each was independently assessed by multiple raters (maximum 59, mean 18). Raters indicated whether they would respond and flagged salient characteristics, requests for action, approaching deadlines, spam signals, or relevance to ongoing projects and meetings. Their annotations were consolidated into seven label categories: "action," "information," "project," "meeting," "spam," "deadline," and "other" (Table 1). We chose this dataset because it uniquely pairs behavioral response data with human-identified perceptual features.

Dataset 2 was drawn from an independent lab study (Singh et al., 2020) and reserved entirely for testing. It records 3,840 individual decisions by 48 participants across 241 distinct emails (roughly 80 per person). The email pool comprises 55 legitimate messages and 186 phishing attempts, none of which overlap with Dataset 1. Because this dataset lacks human-assigned feature labels, it provides a clean test of whether the model generalizes to novel email content.

Table 1: Survey questions presented to end-users during each trial (Xu et al., 2021).

Survey	Summary
Relevance to the recipient	Importance
Request from the recipient to complete a task (e.g., following a link, downloading a file)	Action
Requests a response or input (e.g., replying, sharing documents)	Information
Progress update on a current project or task	Project
Proposes a meeting or conversation	Meeting
Notification of an approaching event or due date	Deadline
Unsolicited, promotional, or suspicious content	Spam
Other	Other

Feature Extraction and Label Aggregation

Using the LUPI framework, we separate test-available inputs from training-only signals. Email text which was accessible at both training and test time, was encoded into dense vectors using the all-MiniLM-L6-v2 sentence-transformer model (Wang et al., 2020; Devlin et al., 2018) via the SentenceTransformer library, producing embeddings that summarize each email's semantic content.

The privileged information consists of survey annotations collected from laboratory participants, capturing their subjective interpretations of each email information that would be absent in a deployed system. To obtain

a single consensus per email, we applied majority voting across all raters, yielding a binary attribute vector over the seven survey categories. Pairing each email's embedding with its consensus label vector frames the learning task as multi-label classification within the IBL pipeline.

Model Architecture

We built a feedforward neural network in Keras (Chollet, 2015) for multi-label prediction where each email may carry several perception tags simultaneously, drawn from participants' survey annotations. A sigmoid output layer produces independent probability estimates for every label category, enabling the network to assign any combination of tags to a single email rather than selecting exactly one. The architecture comprises three layers:

- **Input layer:** 64 units with ReLU activation, accepting vectors whose dimensionality matches the sentence-transformer output; the nonlinearity enables the network to capture complex feature interactions.
- **Hidden layer:** 32 ReLU-activated units that further distill the learned representation into a more compact feature space.
- **Output layer:** one sigmoid unit per survey category, yielding per-label probabilities suited to the multi-label setting.

Optimization used Adam with a learning rate of 0.001, binary cross-entropy loss, and a custom "fuzzy accuracy" metric that counts a prediction as correct when it falls within 0.3 of the target. This tolerance accommodates the inter-rater variability inherent in subjective labeling: different participants often annotate the same email differently, so a strict binary match would understate true model quality. The fuzzy metric thus offers a more faithful performance estimate, consistent with findings that rigid accuracy can be misleading in binary classification and that tolerance-aware evaluation better reflects classifier quality (Singh and Khim, 2022).

Neural Network Model Training and Evaluation

Training runs for up to 100 epochs (batch size 32), with an early-stopping callback that tracks validation loss and halts optimization after 10 consecutive epochs without improvement. When triggered, the network reverts to the weight snapshot that achieved the lowest validation loss, balancing fit quality against generalization risk.

Cognitive Model

The trained network converts each email into a compact predicted-perception vector, which then serves as the instance representation fed to the cognitive model replacing the raw, high-dimensional text embeddings used in earlier work. We instantiate one IBL agent per participant; each agent's memory is seeded with that participant's subset of emails, the associated predicted-perception vectors, and the participant's actual response decisions.

At inference time, a new email is first passed through the neural network to obtain its predicted-perception vector. The IBL agent then retrieves stored instances whose perception vectors are most similar, blends their associated utilities, and selects the action (respond or ignore) with the highest expected value. By operating in this low-dimensional perception space, the agent sidesteps the need to process full email text directly. Figure 1 outlines the framework.

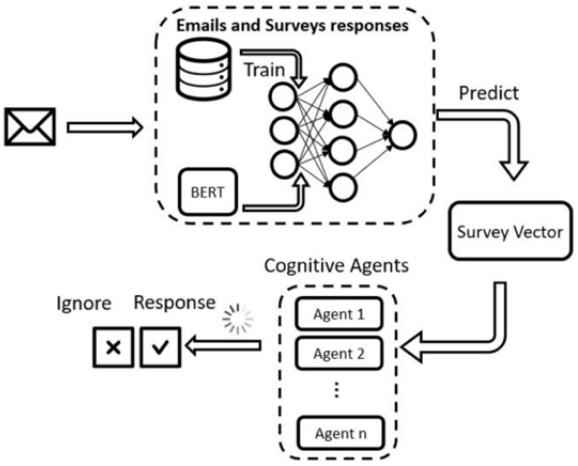

Figure 1: Procedure.

Agents Pooled Decision

We trained one IBL agent per human rater, yielding 84 agents in total. Each agent absorbed the email history and response decisions of a single participant, thereby encoding that individual's judgment tendencies. At test time, every agent independently produced a response prediction for each new email, drawing on its stored memory instances together with the perception vectors output by the neural network. The ensemble decision was reached through majority voting, merging diverse individual viewpoints into one collective prediction. Retaining all 84 agents, even those whose raters assessed fewer emails, preserves the full spectrum of human behaviors in the dataset. This combination of neural-network-generated perception vectors, individualized IBL agents, and pooled output yields a cognitively motivated framework for simulating how people evaluate and react to phishing and legitimate emails alike.

Baseline Model

As a point of comparison, we adopted the BERT-only cognitive model introduced by Xu et al. (2022), which feeds raw sentence embeddings without any human-perception features directly into IBL agents to forecast email responses.

In that baseline, Sentence-BERT (nli-distilroberta-base-v2; Reimers and Gurevych, 2019) produced fixed-length vectors serving as each email's sole instance representation. The IBL agent matched incoming embeddings against

stored memories to select an action, but these vectors encode surface-level semantics without the higher-order, human-salient cues that our privileged-learning pipeline provides. On average, this baseline predicted human responses to phishing emails with less than 60% accuracy. Although BERT excels on many NLP benchmarks, relying exclusively on raw embeddings for phishing-decision modeling proved insufficient to capture the contextual factors that shape human judgment.

RESULTS

Neural Network Performance

Under the privileged-learning setup, the network learned to map email embeddings to perception vectors. We tracked loss and accuracy on both training and validation splits throughout optimization. Early stopping terminated training at epoch 40, the point beyond which validation loss plateaued, confirming that the network had converged without overfitting (Figure 2).

By the final active epoch, classification accuracy stood at 86.97% on the training partition and 87.09% on the held-out validation partition, indicating that the network reliably predicts perception vectors for both familiar and novel emails.

The near-identical figures across splits suggest that the learned mapping transfers well beyond the training set. Overall, the network proved effective at compressing high-dimensional email embeddings into compact perception vectors, a transformation that is critical for enabling cognitive models to operate in a manageable feature space. The early-stopping mechanism further ensured robustness by preventing overfitting.

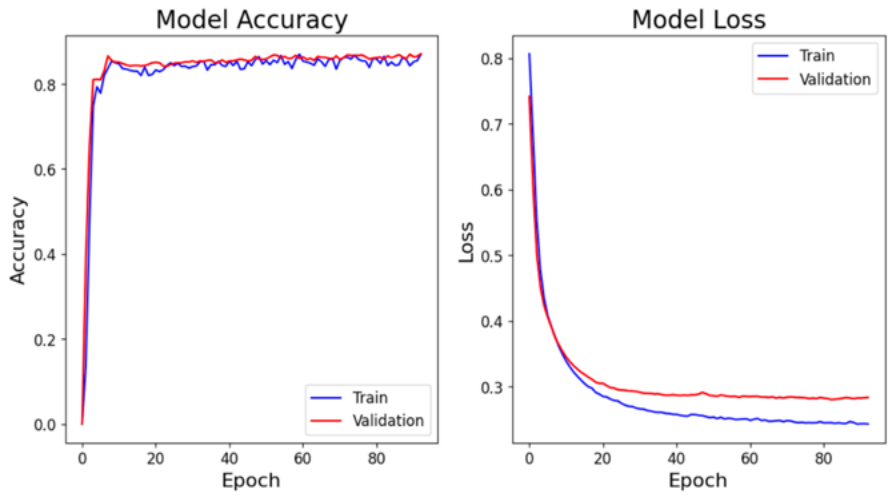

Figure 2: Model performance metrics on two datasets.

Cognitive Agent Performance

Agent performance was quantified with four standard metrics—accuracy, precision, recall, and F1. Because each agent was fitted to one participant's

response history, the evaluation captures how well individualized models reproduce personal decision tendencies. During prediction, agents drew on their stored memory instances together with the neural-network-generated perception vectors, which replaced full email text, showing that compact, perception-aligned representations can support human-like decision modeling.

We measured performance on two fronts: (1) a held-out 20% slice of Dataset 1, testing prediction on emails from the same distribution used in training, and (2) the entirety of Dataset 2, probing whether agents generalize to a wholly different email corpus.

On the first dataset, the agents achieved 73% accuracy, 63.1% precision, 67% recall, and an F1 score of 65%. On the second dataset, which contained only unseen emails, performance reached 62% accuracy, 58% precision, 60% recall, and an F1 score of 59%. These results show that the agents, trained on individualized participant data, can approximate human decision-making patterns with reasonable accuracy and maintain performance when applied to new emails. This demonstrates the effectiveness of using neural-network-derived survey features to model human responses to both phishing and benign emails.

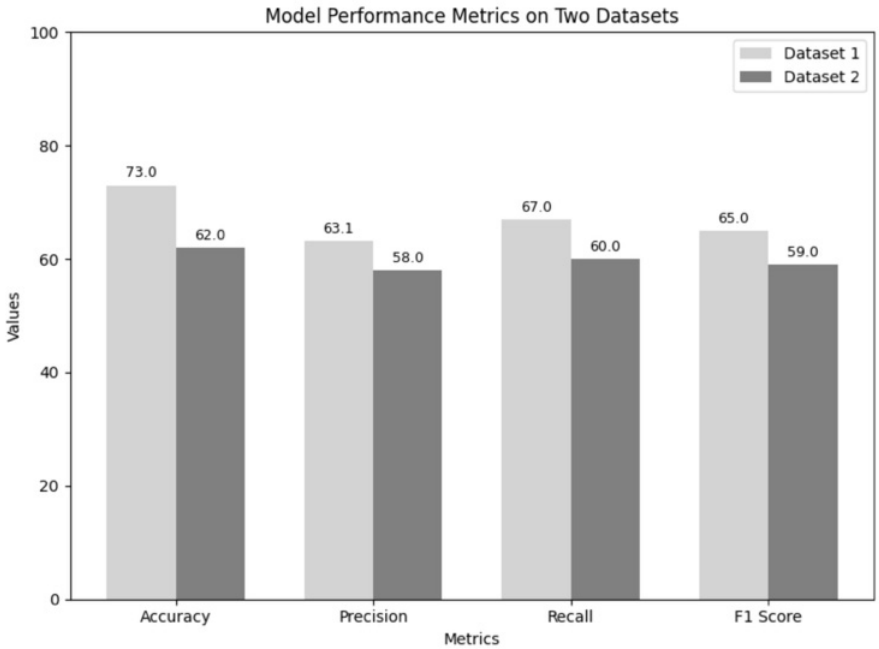

Figure 3: Model performance metrics on two datasets.

DISCUSSION

This work presents a method for modeling email assessment that merges cognitive modeling with machine-learning-based instance engineering. Its central contribution is demonstrating that privileged learning can produce low-dimensional, high-level contextual features aligned with the factors

humans find salient when judging emails which may not be available in production systems but are essential to their performance. Such survey-derived representations sidestep the limitations of hand-coded features (which do not scale) and raw text embeddings (which poorly mirror human encoding and can impair cognitive model accuracy; Xu et al., 2022). Incorporating these human-relevant vectors into cognitive models yields a more cognitively faithful representation of email context and facilitates the processing of lengthy email text within cognitive architectures.

Taken together, the results underscore the benefit of coupling neural feature extraction with cognitively grounded decision architectures. The neural network reliably inferred human-labeled email features from text embeddings, and these inferred features yielded more effective instance representations than full email embeddings. Importantly, our results show that human-salient contextual features can be reliably inferred for unseen emails, enabling scalable and automated instance construction in operational settings and offering a more cognitively grounded account of how people perceive and recall email content.

Performance on the primary dataset shows that cognitive agents can approximate user responses using only predicted survey vectors, indicating that compact, human-aligned representations can effectively replace full email text in cognitive models. However, performance declined on a second dataset of unseen emails, highlighting limits in generalization. This decline likely reflects constraints of the label space: survey-derived labels capture a narrow set of intentions (e.g., information requests or deadlines) while excluding other influential cues such as sender characteristics or required actions, limiting cross-dataset generalization.

A key limitation of this work is the survey-based labeling scheme, which reflects a specific experimental context and may not capture the full range of cues influencing email judgments. As a result, these labels may not generalize to datasets with different structures or content. Additionally, this study did not include an ablation analysis to evaluate the individual contribution of each survey-derived feature. The seven survey categories used here (e.g., action, information, meeting) were drawn from a single experimental context and treated as a unified intent/content representation rather than as independently motivated dimensions. As a result, their individual predictive value was not assessed. In follow-up work, we address this gap by grounding feature selection in a broader literature-based taxonomy of email intent categories, enabling systematic evaluation of which perceptual dimensions contribute most to cognitive model performance. Future work should also explore how richer and more diverse feature sets, covering sender characteristics, urgency cues, and rhetorical strategies can further improve model fidelity and cross-dataset generalization.

Another limitation is the reliance on lab-based datasets, which, while useful for controlled evaluation, do not capture the full diversity and complexity of real-world phishing emails. Real-world email environments involve greater variation in sender reputation, message sophistication, and user context, all of which may influence how individuals encode and respond to threats. Validating the proposed framework on larger, naturalistic datasets, ideally

drawn from organizational email logs remains an important direction for future work and would provide a stronger test of the model's robustness across diverse threat landscapes.

Overall, this study advances cognitive modeling of phishing and benign email evaluation by demonstrating privileged learning as a practical approach for constructing low-dimensional, human-relevant feature vectors. Acknowledging current limitations situates this work within a broader effort toward scalable and generalizable frameworks. Grounding email evaluation in human-centered cognitive features provides a foundation for future research on user decision making and the design of more effective, tailored anti-phishing training strategies.

ACKNOWLEDGMENT

This work was supported by a grant from the National Science Foundation (NSF grant # 2142888). The opinions, findings, and conclusions do not reflect the views of the funding agencies, cooperating institutions, or other individuals.

REFERENCES

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review*, *111*(4), 1036.

Azuma, R., Daily, M., & Furmanski, C. (2006). A review of time critical decision making models and human cognitive processes. In *2006 ieee aerospace conference*.

Chollet, F. (2015). *Keras: Deep learning library for python*. (GitHub Repository, Retrieved from https://github.com/fchollet/keras)

Cranford, E. A., Lebiere, C., Rajivan, P., Aggarwal, P., & Gonzalez, C. (2019). Modeling cognitive dynamics in enduser response to phishing emails. In *Proceedings of the 17th iccm*.

Cranford, E. A., Singh, K., Aggarwal, P., Lebiere, C., & Gonzalez, C. (2021). Modeling phishing susceptibility as decisions from experience. In *Proceedings of the 17th international conference on cognitive modeling (iccm)*.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint*.

Dhamija, J. D., Tygar, J. D., & Hearst, M. (2006). Why phishing works. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 581–590).

Ferreira, A., & Teles, S. (2019). Persuasion: How phishing emails can influence users and bypass security measures. *International Journal of Human-Computer Studies*, *125*, 19–31.

Gonzalez, C., Lerch, J. F., & Lebiere, C. (2003). Instance-based learning in dynamic decision making. *Cognitive Science*, *27*(4), 591–635.

Gonzalez, C., Ben-Asher, N., Oltramari, A., & Lebiere, C. (2014). Cognition and technology. In Cyber defense and situational awareness (pp. 93–117). Springer.

Hakim, Z. M., Ebner, N. C., Oliveira, D. S., Getz, S. J., Levin, B. E., Lin, T., & Wilson, R. C. (2021). The phishing email suspicion test (pest): A lab-based task for evaluating the cognitive mechanisms of phishing detection. *Behavior Research Methods*, *53*, 1342–1352.

Ho, G., Mirian, A., Luo, E., Tong, K., Lee, E., Liu, L., ... & Voelker, G. M. (2025, May). Understanding the efficacy of phishing training in practice. In *2025 IEEE Symposium on Security and Privacy (SP)* (pp. 37–54). IEEE.

Kumaraguru, P., Rhee, Y., Acquisti, A., Cranor, L. F., Hong, J., & Nunge, E. (2007). Protecting people from phishing: The design and evaluation of an embedded training email system. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 905–914).

Kumaraguru, P., Sheng, S., Acquisti, A., Cranor, L. F., & Hong, J. (2010). Teaching johnny not to fall for phish. *ACM Transactions on Internet Technology (TOIT)*, *10*(2),1–31.

Malloy, T., & Gonzalez, C. (2024). Applying generative artificial intelligence to cognitive models of decision making, *Frontiers in Psychology*, *15*, 1387948.

Mehrabi, E., and Rajivan, P. (2025). Neural network-driven cognitive models of phishing decisions: Evaluating a privileged learning framework for instance representation. In Proceedings of the 23rd International Conference on Cognitive Modeling (ICCM). [Poster presentation].

Mikolov, T., Karafiát, M., Burget, L., Cernocký, J., & Khudanpur, S. (2010). Recurrent neural network based language model. In *Interspeech* (Vol. 2, pp. 1045–1048).

Momeni, A., Tatwawadi, K., & Stanford, U. (2018). Understanding LUPI (learning using privileged information). *Ionosphere*, *201*(7), 6.Morrison, D., & Gonzalez, C. (2024). *PyIBL: Python implementation of Instance-Based Learning* [Software documentation]. Dynamic Decision Making Laboratory. https://ddm-lab.github.io/pyibl-documentation

Newell, A., & Simon, H. (1972). *Human problem solving*. Oxford, England: Prentice-Hall.

Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint*. Sannigrahi, S., Genabith, J. V., & España-Bonet, C. (2023). Are the best multilingual document embeddings simply based on sentence embeddings? *arXiv preprint*.

Sawyer, B. D., & Hancock, P. A. (2018). Hacking the human: The prevalence paradox in cybersecurity. *Human Factors*, *60*(5), 597–609.

Shonman, M., Shi, X., Kang, M., Wang, Z., Li, X., & Dahbura, A. (2022). Using a computational cognitive model to understand phishing classification decisions. In *35th international bcs human-computer interaction conference* (pp. 1–10).

Singh, K., Aggarwal, P., Rajivan, P., & Gonzalez, C. (2020, December). What makes phishing emails hard for humans to detect? In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 64, pp. 431–435). Los Angeles, CA: SAGE Publications.

Singh, S., & Khim, J. T. (2022). Optimal binary classification beyond accuracy. In *Advances in neural information processing systems 35* (pp. 18226–18240).

Vapnik, V., & Vashist, A. (2009). A new learning paradigm: Learning using privileged information. *Neural Networks*, *22*(5), 544–557.

Vishwanath, A., Harrison, B., & Ng, Y. J. (2018). Suspicion, cognition, and automaticity model of phishing susceptibility. *Communication Research*, *45*(8), 1146–1166.

Wang, W., Bi, B., Yan, M., Wu, C., Xia, J., Li, Z., ... Si, L. (2020). Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *arXiv preprint*.

Wu, M., Miller, R. C., & Garfinkel, S. L. (2006). Do security toolbars actually prevent phishing attacks? In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 601–610).

Xu, T., Singh, K., & Rajivan, P. (2021). Spearsim: Design and evaluation of synthetic task environment for studies on spear phishing attacks. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 65, pp. 1500– 1504). SAGE Publications Sage CA: Los Angeles, CA.

Xu, T., Singh, K., & Rajivan, P. (2022). Modeling phishing decision using instance based learning and natural language processing. In *Proceedings of the hawaii international conference on system sciences (hicss)* (pp. 1–10).