

©Copyright 2025

Elaheh Mehrabi

Evaluating text representations in cognitive models of human decision
making: The case of phishing attacks

Elaheh Mehrabi

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2025

Reading Committee:

Prashanth Rajivan, Chair

Shuai Huang

Ashis Banerjee

Program Authorized to Offer Degree:
Industrial and System Engineering

University of Washington

Abstract

Evaluating text representations in cognitive models of human decision making: The case of phishing attacks

Elaheh Mehrabi

Chair of the Supervisory Committee:
Prashanth Rajivan
Department of Industrial and Systems Engineering

Phishing attacks use deceptive emails to manipulate individuals into revealing sensitive information or performing harmful actions. They continue to remain a widespread threat because they exploit human cognitive vulnerabilities rather than technological weaknesses. Unlike hardware or software weaknesses that can often be patched, human susceptibility to phishing cannot be intervened with a one-size-fits-all anti-phishing training solution. People differ in the types of phishing threats to which they are susceptible, which require personalized and adaptive anti-phishing training solutions capable of tailoring learning experiences to individual needs and context.

Cognitive models could enable analyses of cognitive processes predictive of individuals' susceptibilities to phishing threats. They could enable anti-phishing training solutions to identify gaps in individuals' knowledge about phishing threats, assess people's ability to recall training when faced with real threats, and estimate how well individuals can generalize training received to recognize new types of attacks. However, a key challenge in the developing such cognitive models lies with *instance representation*. In the context of models of phishing decision making, it would be the information communicated through email text. It remains unclear how people encode and recall past email conversations when making decisions about phishing attacks. This dissertation addresses this critical gap by investigating how different text representations affect cognitive models of human decision-

making in phishing contexts. Specifically, it explores how people encode and recall textual information from emails when deciding whether to engage with or ignore them, and how to generate representations aligned with human memory processes for use in computational cognitive models.

My research began by integrating existing cognitive models into a simulated environment emulating user interactions with phishing emails. Early experiments revealed that these models struggled to capture the processes by which humans recall and interpret complex email content when making response decisions, highlighting a fundamental limitation in current modeling approaches.

As a first step toward addressing this gap, I developed a privileged learning framework that was trained on a set of high-level cues derived from human-labeled data. Once trained, the model was able to generate similar representation cues for new, unseen emails that lacked these annotations. This approach demonstrated the potential to bridge the gap between raw email text and more abstract decision-making features, while also highlighting the need for deeper analysis of how different representation methods impact cognitive modeling.

Motivated by these findings, I systematically evaluated two broad categories of text representations within the instance-based learning (IBL) framework. The first category included lower-order, text-based representations, such as embeddings of full sentences, embeddings of email summaries from generative AI, and recall-based keywords. The second category encompassed higher-order, intent-driven representations capturing cues like tone, sentiment, and inferred communicative intent—features inspired by speech act theory hypothesized to better reflect how humans remember the intent behind what was communicated through an email. Through extensive evaluation within the instance-based learning theory framework, my findings demonstrate that higher-level, abstract representations—such as inferred intent, sentiment, and tone—more accurately model human decision-making than full-text embeddings or even embeddings based on email summaries. These representations not only improved prediction performance but also enhanced robustness when simulating users whose behaviors were less consistent or dissimilar from the majority, suggesting greater

generalizability across diverse individuals.

In a subsequent exploratory effort, I implemented IF-THEN production rules based on intent representations within the ACT-R cognitive architecture to examine how rule-based cognitive modeling could enhance model alignment with human responses. While preliminary, this effort highlighted the potential of explicit rule-guided frameworks to support future human decision-making modeling directions and offer interpretable decision mechanisms.

Together, these contributions advance our understanding of how cognitively grounded representations of email content influence the modeling and analysis of phishing decisions. They underscore the value of aligning computational models with human memory processes to support the development of personalized, adaptive, and cognitively-informed anti-phishing interventions.

TABLE OF CONTENTS

	Page
List of Figures	iii
Glossary	v
Chapter 1: Introduction	1
1.1 Research Questions and Objectives	3
1.2 Structure	4
Chapter 2: Literature Review	7
2.1 Phishing and Spear Phishing:	7
2.2 Instance-Based Learning:	8
2.3 Human Memory	11
2.4 Natural Language Processing (NLP)	13
Chapter 3: SpearSim-v2: Experimental Design and Attacker Behavior Analysis . .	16
3.1 Introduction	16
3.2 Objective	17
3.3 Method	18
3.4 Results	23
3.5 Discussion	25
3.6 Limitations and Future Work	26
Chapter 4: Privileged Learning for Modeling Human Decision-Making in Phishing Attacks	27
4.1 Introduction	27
4.2 Objective	28
4.3 Method	28
4.4 Results	34
4.5 Discussion	38

Chapter 5:	Instance representations through intent modeling	41
5.1	Representation Strategies for Modeling Human Recall	41
5.2	Intent Modeling	42
5.3	Data Collection:	45
5.4	Method	55
5.5	Evaluation	57
5.6	Results	58
5.7	Overall Performance Comparison	61
5.8	Individual Agent Analysis	66
5.9	Prediction Agreement by Email Type	69
5.10	Discussion and Future Work	71
Chapter 6:	Exploring ACT-R based phishing decision making models	74
6.1	Motivation	74
6.2	Introduction and Background	75
6.3	Objectives	77
6.4	Method	78
6.5	Discussion	88
Chapter 7:	Conclusion	90
7.1	Contributions	92
7.2	Future Directions	93
7.3	Closing Summary	93
Bibliography		94
Appendix A:	Survey Questionnaire Items	107

LIST OF FIGURES

Figure Number		Page
1.1	Overview of Human Memory Activation and Retrieval in Phishing Response: Connections between Dissertation Topics	5
3.1	Integrating Software and the Cognitive Model Procedure Overview	19
3.2	Attacker Interface with an example attack	20
3.3	Attacker interface after attack deployment	20
3.4	Possible Responses provided by the software	21
3.5	Software and Human Response Time	24
4.1	Procedure	32
4.2	Model Accuracy and Loss for training and validation	35
4.3	Model Performance Metrics on Two Datasets	37
5.1	Hierarchy of email representations ranging from lower-level to higher-level abstract representations	45
5.2	The most common recalled words across all the emails	53
5.3	Distribution of chosen intents and sentiment labels across emails	55
5.4	Distribution of email types (Phishing, Primary, Promotion) across individual raters.	60
5.5	Average log-likelihood per agent for the Summary-Based Model.	63
5.6	Average log-likelihood per agent for the Keyword-Based Model.	63
5.7	Average log-likelihood per agent for the Intent-Based Model.	63
5.8	Average log-likelihood per agent for the Intent + Sentiment + Tone Model. . .	64
5.9	PCA projection of participant performance profiles. Clusters represent dis- tinct groups of raters based on model fit metrics.	65
5.10	Average F1 scores by cluster for each model.	66
5.11	Log-likelihood heatmaps for an agent from Cluster 1 (inconsistent rater). . .	67
5.12	Log-likelihood heatmaps for an agent from Cluster 0 (consistent rater). . . .	68
5.13	Prediction agreement for the Summary-Based model across email types. . . .	69
5.14	Prediction agreement for the Keyword-Based across email types.	69
5.15	Prediction agreement for the Intent-Only model across email types.	70

5.16	Prediction agreement for the Intent-Sentiment-Tone model across email types.	70
6.1	Example of a matched rule firing directly.	87
6.2	Example of retrieval-based decision due to no matching rule.	87

GLOSSARY

NLP: Natural Language Processing

LLM: Large Language Model

IBL: Instance-based Learning

ACKNOWLEDGMENTS

This dissertation marks the culmination of a journey filled with learning, perseverance, and personal growth. I am profoundly grateful to the many people who made this accomplishment possible, both personally and professionally.

First and foremost, I would like to express my deepest appreciation to my advisor, Prof. Prashanth Rajivan. His thoughtful guidance, patient mentorship, and unwavering support have been instrumental at every stage of this work. Prof. Rajivan's ability to challenge me to think critically while offering steady encouragement helped me grow immensely as a researcher. His insightful feedback, commitment to scientific rigor, and dedication to fostering my development as an independent thinker have left a lasting impact on both my academic journey and my personal growth. I am deeply fortunate to have had the privilege of working under his mentorship, and I will carry the lessons I learned from him far beyond this dissertation.

I am also sincerely grateful to my dissertation committee members, Prof. Shuai Huang, Prof. Ashis Banerjee, and Prof. Andrea Stocco, for their invaluable feedback, thoughtful questions, and guidance. Their diverse expertise and perspectives helped shape this work and pushed me to refine my ideas in meaningful and productive ways.

I am thankful for the opportunity to be part of Prof. Rajivan's lab, where I found not only an intellectually stimulating environment but also a truly collaborative and supportive one. Working alongside talented and dedicated labmates has been one of the most rewarding parts of this journey. The countless discussions, brainstorming sessions, and shared challenges have enriched my experience in ways I will always treasure.

I also extend my gratitude to the University of Washington, which has been a home for my academic and professional growth over the course of this program. The faculty, staff, and peers at UW provided an environment where ideas could flourish and where I was

constantly encouraged to push the boundaries of my research. I am deeply appreciative of the opportunities and resources that made this work possible.

Heartfelt gratitude goes to my family. To my mother, whose endless love, patience, and compassion have been my anchor. Her constant check-ins, her comforting words when things felt overwhelming, and her ability to celebrate even my smallest wins gave me the strength to keep moving forward. To my father, whose unwavering belief in me has been a guiding light. His encouragement, wise advice, and powerful support have given me the confidence to take on challenges I once thought impossible. Knowing he has always been in my corner has made all the difference. To my sister, Atefeh, for being more than just a sibling — for being a mentor, role model, and constant source of inspiration. Having already walked this path, her thoughtful guidance and genuine understanding of the struggles of this journey motivated me to persevere, even during the most difficult moments. Together, their love and encouragement turned every obstacle into an opportunity and made this achievement deeply meaningful. Without their unwavering support, this journey would not have been possible.

DEDICATION

To my parents, Mehran and Goli, and my sister, Atefeh.

Chapter 1

INTRODUCTION

In today’s information-rich digital landscape, people must make several decisions based on the information conveyed through unstructured text. From online advertisements to social media posts and emails, many forms of communication on the Internet are primarily text, and depending on how people process and recall such information from memory, can significantly influence their decisions and behaviors, sometimes with serious consequences. The challenge of modeling and analyzing human decisions in such contexts has become crucial, particularly in areas such as cybersecurity, where understanding decision-making processes can directly inform defense and training intervention strategies [132]. One context where these concerns are particularly relevant is phishing.

Phishing remains a critical security threat despite decades of research because it targets vulnerabilities in human cognition, as opposed to technological weaknesses [46]. Phishing attacks employ psychological manipulation to influence individuals into revealing sensitive personal information, such as login credentials or financial data [7]. These attacks frequently impersonate trusted sources, such as banks, employers, popular technology brands, or government agencies, making them particularly convincing and difficult to detect.

Falling victim to phishing can have devastating consequences for individuals and organizations alike. According to the FBI’s Internet Crime Complaint Center (IC3) 2024 Annual Report, Americans reported losses totaling \$16.6 billion in 2024 alone, marking the highest annual losses ever recorded by IC3 and reflecting a significant increase over previous years [43]. Over the past five years, losses reported to IC3 have exceeded \$50 billion, illustrating the increasing financial impact of phishing and related cyber-enabled fraud. These rising losses underscore the urgent need for improved defenses and training solutions that address the human vulnerabilities exploited by phishing attacks.

Spear phishing, a more targeted form of phishing, intensifies this challenge by using

personalized information about recipients to craft convincing messages, greatly increasing the likelihood of success [59]. The sophistication of these emails makes them difficult for detection algorithms to identify, as traditional systems often rely on generic indicators found in mass phishing campaigns[57]. Spear phishing poses significant challenges for human users, who may struggle to distinguish carefully tailored malicious emails from legitimate communications. These factors underscore the urgent need for novel, human-aware anti-phishing training strategies that incorporate models of human cognition and real-world decision-making processes.

One approach to addressing this challenge is the development of anti-phishing training solutions capable of producing personalized training experiences using models predictive of individual susceptibilities to phishing threats. Such solutions could help identify gaps in an individual’s awareness, assess their ability to recall different varieties of phishing threats, assess how well they remember training when faced with real attacks, and determine their ability to generalize that training to recognize new forms of phishing. Developing these advanced training systems requires valid models of human response to phishing threats. Models grounded in cognitive architectures, such as ACT-R and Instance-Based Learning Theory (IBLT) [48], are particularly attractive because they can be used to analyze, simulate, and predict individuals’ susceptibility to various types of phishing attacks [34, 144].

However, a key challenge in the development of such cognitive models lies with *instance representation*. In the context of models of phishing decision-making, it would be the information communicated through email text. It remains unclear how people encode and recall past email conversations (including other phishing emails and phishing training) when making decisions about phishing attacks. Real-world emails are context-rich and multilayered, and people may not store or recall such conversations at the text level. Instead, their memory representations likely emphasize latent structures such as communicative intent, topic, tone, and entities involved in the conversation. Therefore, this dissertation aims to address a key research question: how to effectively represent long text emails within cognitive models of human decision-making. To explore this, I combine behavioral experiments with cognitively grounded models based on IBL principles, focusing on ways to abstract and represent email content in formats that align more closely with human memory and

reasoning.

Traditional approaches to representing text in cognitive models often rely on basic surface-level features, such as words and full sentence embeddings, which may not fully capture the complex contextual meanings that influence human decision-making [144, 34]. As an initial attempt to find alternatives to modeling full unstructured text, I experimented with using survey-derived human responses presented as discrete features or “chunks” representing key aspects of emails. While this privileged approach demonstrated promising predictive performance, it also revealed limitations in interpretability and generalizability, as the fixed and limited set of features made it difficult to comprehensively explain human thought processes. Motivated by these observations, my research shifted to a systematic investigation of email representations from two main perspectives: (1) lower-level, text-based features such as full-text embeddings, summaries, and keywords, and (2) higher-level, abstract semantic features capturing inferred intent, sentiment, and tone. This comparative analysis aimed to determine which type of representation better aligns with human decision-making processes when responding to phishing emails.

1.1 Research Questions and Objectives

The primary objective of this dissertation is to identify effective representations of long, complex emails within cognitive models of human decision-making, with the aim of advancing our understanding of how different types of features align with human processing and recall mechanisms when making decisions about phishing emails. Modeling human responses to phishing emails is uniquely challenging because instance-based learning (IBL) models, which rely on similarity-based retrieval from memory, work best with higher-order, semantically meaningful features that create a coherent memory space. In contrast, raw, unstructured emails are long, variable in length, and context-rich, making it unclear how individuals encode, recall, and reason over such complex information. This creates a significant gap in our ability to develop cognitively grounded models that realistically simulate human email processing and phishing susceptibility.

This research began with a technical goal: integrating existing IBL models into a phishing simulation to automate user responses. However, during this process, it became clear

that these models did not fully capture how people recall, interpret, and evaluate phishing emails. This disconnect led to a reframing of the research toward understanding how to represent long-form, unstructured emails in a way that aligns with human cognitive processing. Building on this realization, the focus shifted from merely embedding existing models into simulations to systematically investigating which types of text features and cognitive modeling strategies best reflect human decision-making in phishing contexts.

To address these challenges, this dissertation investigates the following research questions.

1. Can we learn high-level cognitive cues from labeled data and generalize them to new, unlabeled emails?
2. Which types of text representations better align with human recall and decision-making?
3. Can rule-based cognitive models (like ACT-R) effectively simulate phishing decisions using learned cues?

The efforts in this dissertation contribute to a comprehensive investigation of how different text representations and modeling strategies influence the ability of cognitive models to simulate human responses to phishing emails. Figure 1.1 illustrates the research flow of this dissertation.

1.2 Structure

This dissertation is organized into seven chapters, each addressing key components of the research.

Chapter 2 presents a comprehensive review of the literature on phishing attacks and cognitive modeling. It covers social engineering techniques, psychological factors contributing to phishing susceptibility, and prior approaches leveraging natural language processing and machine learning, with an emphasis on the gaps in modeling human decision-making processes.

Chapter 3 introduces the development of an automated simulation environment for phishing decision-making studies. This environment allowed for the integration and testing of cognitive models in simulated user interactions.

Chapter 4 serves as a starting point by leveraging an existing annotated dataset in which participants categorized each email by topic. This chapter explores how these survey-based annotations can be used to represent emails in cognitive models as an alternative to raw text, providing an initial step toward building more cognitively meaningful representations.

Chapter 5 then systematically compares alternative text representations, ranging from lower-level representations such as summaries and keywords to higher-level, abstract semantic features including inferred intent, sentiment, and tone, to evaluate their alignment with human recall and decision-making.

Chapter 6 explores rule-based cognitive modeling using ACT-R production rules as a complementary approach, examining whether interpretable, rule-driven models can capture decision-making patterns observed in human participants. Figure 1.1 illustrates the research flow of this dissertation and how it connects to human memory and activation processes.

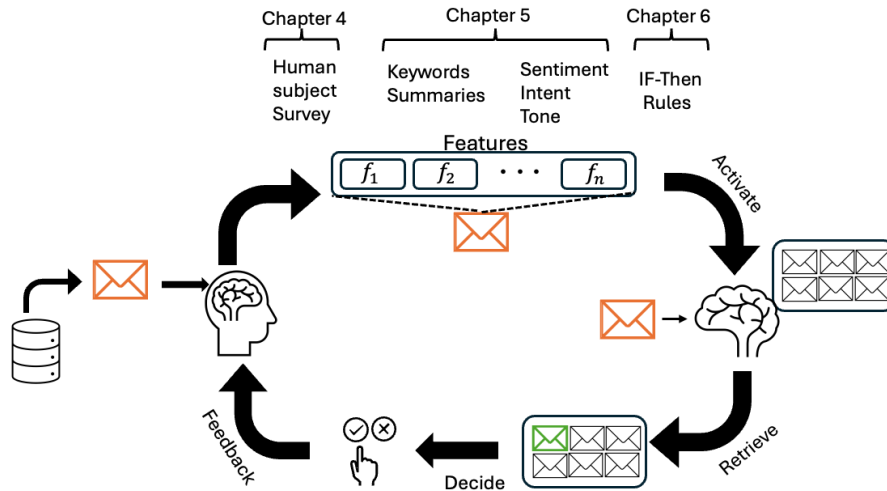


Figure 1.1: Overview of Human Memory Activation and Retrieval in Phishing Response: Connections between Dissertation Topics

Together, these investigations aim to advance our understanding of effective text representations and modeling strategies for simulating human decision-making when evaluating phishing emails.

Finally, Chapter 7 provides a discussion of the findings, synthesizing insights from all chapters, highlighting contributions to the field, and outlining directions for future research to further advance cognitively informed phishing defenses.

Chapter 2

LITERATURE REVIEW

2.1 *Phishing and Spear Phishing:*

Phishing has emerged as a significant threat in the digital landscape, characterized by deceptive practices aimed at acquiring sensitive information such as usernames, passwords, and credit card details [2]. Phishing attacks usually involve fake emails and websites that pretend to be from trusted sources, tricking people into giving away private information, or taking insecure actions like downloading malicious payloads [58]. The success of these attacks relies heavily on the attackers' ability to exploit human cognitive biases and weaknesses. In contrast, spear phishing targets specific individuals or organizations, making it a more personalized and dangerous variant of phishing. Spear phishing campaigns are often meticulously designed, utilizing the information collected about the target to enhance credibility and increase the likelihood of success [13]. This tailored approach not only improves the effectiveness of the attack, but also complicates detection efforts for potential victims[56, 87, 17]. Research has indicated that spear phishing poses a greater risk to organizations, as successful attacks can lead to significant financial losses, data breaches, and reputational damage [49].

Understanding the psychological mechanisms that underpin susceptibility to phishing and spear phishing is essential in developing effective countermeasures [47, 134]. The rapid evolution of phishing techniques requires ongoing research on user behavior and decision-making processes to enhance security awareness and resilience against such threats. The intersection of phishing tactics and cognitive mechanisms highlights the need for a deeper understanding of how people process information and make decisions in response to phishing attempts [1, 129, 138].

The principles of Instance-Based Learning (IBL) could offer valuable insights into this phenomenon, suggesting how people rely on past experiences to inform their actions [48].

By exploring IBL, we can better understand the cognitive underpinnings of phishing vulnerability and develop strategies that leverage these insights to mitigate risks [144].

2.2 Instance-Based Learning:

Instance-based learning models describe decision-making as a process of retrieving and generalizing from past experiences, or instances, that resemble the current situation. Each instance is stored in memory as a representation of the task context, the decision taken, and the outcome, enabling the model to adapt relevant prior experiences to guide current choices [48, 144]. This concept is particularly relevant in the context of understanding how individuals respond to phishing and spear phishing attempts.

2.2.1 IBL Theory and Framework

At the core of Instance-Based Learning (IBL) is the notion that memory is organized around specific instances rather than generalized concepts [48]. This framework suggests that learning occurs through the retrieval and evaluation of previously encountered instances in similar contexts. One of the key components of IBL is the Instance-Based Learning Theory (IBLT), which posits that individuals store detailed instances in memory and use them to make predictions or decisions when faced with new situations [48].

A central concept in IBL is **activation**, which refers to the process of retrieving stored instances based on their similarity to the current context, their frequency, and recency. Activation strength reflects how readily an instance can be accessed, depending on factors such as frequency of exposure, recency of past encounters, and contextual similarity to the current situation.

Another essential mechanism is **blending** mechanism. In instance-based learning models, actions are generated using a blending mechanism that draws on past instances. This process retrieves information from multiple memories that share similar contextual features and produces a consensus value, rather than relying on a single matching memory [67, 48]. This blended value aggregates outcomes of relevant past instances, weighted by how likely they are to be retrieved, to generate a **decision utility** for each option. The option with

the highest expected utility—representing the combination of past outcomes and their probabilities—is selected as the preferred choice.

These processes enable rapid, experience-based judgments, which are particularly relevant in tasks like phishing detection, where users must quickly assess whether a message is legitimate or suspicious.

Understanding how these cognitive mechanisms operate in phishing contexts can inform strategies for modeling user responses and designing more effective interventions.

2.2.2 Cognitive Mechanisms in IBL:

Research has demonstrated that IBL is closely tied to various cognitive mechanisms, such as memory retrieval, pattern recognition, and judgment. The retrieval of instances from memory is influenced by factors such as recency and frequency of exposure, highlighting the importance of experience in shaping responses to phishing attempts [88]. Furthermore, individuals often rely on heuristics—mental shortcuts derived from past instances—to make quick judgments about the authenticity of messages, impacting their susceptibility to phishing attacks [45]. In the topic of phishing, the role of IBL underscores the importance of education and training that leverages individuals’ existing experiences. By exposing users to simulated phishing scenarios and highlighting past instances of successful attacks, it is possible to enhance their ability to recognize and react to future threats. This approach aligns with findings that indicate enhanced training methods that emphasize IBL principles can improve phishing detection and reduce vulnerability [88].

Mathematical Formulation of IBL

In Instance-Based Learning, decisions are guided by estimating the expected utility of each option, calculated through the process of blending retrieved past instances. The expected utility for option k at trial t is computed as:

$$V_{k,t} = \sum_{i=1}^n P_{i,k,t} \times X_{i,k,t}, \quad (2.1)$$

where:

- $X_{i,k,t}$ represents the outcome (utility) of instance i for option k at trial t .
- $P_{i,k,t}$ is the probability of retrieving instance i from memory when considering option k at trial t .

The retrieval probability $P_{i,k,t}$ is determined by the relative activation of each instance:

$$P_{i,k,t} = \frac{e^{A_{i,k,t}/\tau}}{\sum_{i=1}^n e^{A_{i,k,t}/\tau}}, \quad (2.2)$$

where τ the softmax temperature parameter, often defined as a function of retrieval noise, which controls the stochasticity of choice based on activation values.

The activation $A_{i,k,t}$ quantifies how accessible an instance is in memory, combining the effects of frequency, recency, contextual similarity, and stochastic noise:

$$A_{i,k,t} = \ln \left(\sum_{t_i=1}^{t-1} (t - t_i)^{-d} \right) + MP \sum_k Sim(v_k, c_k) + \sigma \ln \left(\frac{1 - \gamma_{i,k,t}}{\gamma_{i,k,t}} \right), \quad (2.3)$$

where:

- d is the decay parameter controlling forgetting over time.
- MP is the mismatch penalty scaling the contribution of contextual similarity.
- $Sim(v_k, c_k)$ measures the similarity between the current cue c_k and past instance v_k .
- σ scales the Gaussian noise; $\gamma_{i,k,t}$ is a uniformly distributed random variable in $(0,1)$.

These equations formalize how **activation** influences **retrieval probabilities**, which in turn determine the **blended expected utility** guiding decision-making in IBL. By capturing cognitive factors like recency, similarity, and stochastic variability, this model reflects how individuals leverage past experiences to make decisions under uncertainty — a process highly relevant for understanding susceptibility and responses to phishing attempts.

To effectively model human cognition, it is essential to delve deeper into the underlying processes of human memory. Understanding how memories are formed, stored, and retrieved will enhance our comprehension of decision-making in the context of phishing. This section

will explore the intricacies of human memory, providing insights into its structure and function, which are vital for grasping how individuals navigate and respond to these digital threats.

2.3 Human Memory

Human memory is a complex and multifaceted system that plays a critical role in cognitive processes, including learning, decision-making, and problem-solving. It can be broadly categorized into three main types: sensory memory, short-term memory, and long-term memory, each serving distinct functions in the encoding and retrieval of information [11].

2.3.1 Structure of Memory

Sensory Memory is the initial stage of memory, where information from the environment is briefly stored in its raw form. This type of memory has a very short duration, typically lasting only a few seconds. Sensory memory allows individuals to process and retain information about stimuli in their environment, such as sounds and visual images, for immediate use [118]. Short-term Memory (STM), often referred to as working memory, holds a limited amount of information for a short period. STM is essential for tasks that require active processing, such as reasoning and comprehension. Research has shown that the capacity of short-term memory is typically around seven items [82]. The temporary nature of STM necessitates strategies such as rehearsal to maintain information for extended periods. Long-term Memory (LTM) is a more permanent storage system that can retain vast amounts of information for extended periods, from days to a lifetime. LTM is further divided into explicit (declarative) and implicit (non-declarative) memory. Explicit memory encompasses facts and events that can be consciously recalled, while implicit memory includes skills and conditioned responses that are typically performed unconsciously [125].

2.3.2 Encoding and Retrieval

The processes of encoding and retrieval are fundamental to the functioning of human memory. Encoding refers to the transformation of information into a form that can be stored

in memory. This process is influenced by various factors, including attention, motivation, and the emotional significance of the information [33]. Effective encoding strategies, such as visualization and organization, can enhance the likelihood of successful memory retention.

Retrieval involves accessing stored information when needed. The effectiveness of retrieval can be affected by contextual cues and the way in which the information was initially encoded [126]. The retrieval practice effect suggests that frequent retrieval of information strengthens memory traces, thereby improving long-term retention [102]. Understanding these retrieval mechanisms is particularly relevant when considering how individuals recognize and respond to phishing attempts.

2.3.3 Memory in Context of Phishing

In the context of phishing, the dynamics of human memory play a significant role in susceptibility to attacks. Users' past experiences with similar threats are stored in their long-term memory and can influence their judgment when encountering potential phishing attempts. For example, individuals who have previously fallen victim to phishing may become more cautious and employ strategies learned from past experiences to avoid similar traps in the future [40].

Research has shown that individuals often rely on heuristics—mental shortcuts that draw upon past experiences—to make quick decisions about the authenticity of messages. These heuristics, while efficient, can also lead to errors in judgment, especially when users fail to recognize the subtleties of a sophisticated phishing attack [113]. Therefore, enhancing users' memory through targeted training and education can significantly improve their ability to identify and respond to phishing threats effectively.

2.3.4 The Role of Words in Memory Storage

Words are not just symbols but are integral to our cognitive processes, influencing how information is encoded and retrieved. Port [90] highlights the significance of phonological representations and how they relate to the mental lexicon—our internal store of words and their meanings. This lexicon is critical for understanding language and is closely tied

to memory retrieval. When we encounter written or spoken words, they activate related memories and concepts stored within our cognitive framework. This activation process is essential in distinguishing relevant information, particularly in scenarios like phishing, where familiar linguistic patterns can signal authenticity or deceit.

As we delve deeper into understanding human cognition and memory, it becomes increasingly clear that the processing of textual information plays a vital role in shaping our decision-making capabilities, particularly in contexts like phishing. Given that much of the communication in phishing attempts is text-based, leveraging Natural Language Processing (NLP) tools is essential for analyzing, interpreting, and modeling this data effectively. NLP techniques, such as word embeddings and advanced models like BERT, enable us to capture the semantic meaning of words and phrases, facilitating the construction of cognitive models that can simulate human responses to textual threats. In the next section, I will explore the key NLP methodologies that enhance our ability to analyze text data, drawing connections between linguistic features and cognitive processes.

2.4 Natural Language Processing (NLP)

Natural Language Processing (NLP) encompasses a variety of techniques that enable machines to understand, interpret, and generate human language. As text plays a central role in phishing communications, NLP becomes an indispensable tool in developing models that can analyze and respond to these threats effectively. The application of NLP techniques, particularly in the context of embedding methods and semantic similarity, provides insights into the underlying cognitive processes that influence how individuals perceive and react to phishing messages.

2.4.1 Word Embeddings

Word embeddings are a foundational NLP technique that represents words as dense vectors in a continuous vector space. This representation captures semantic relationships between words, allowing for the measurement of word similarity based on their contextual usage. Techniques such as Word2Vec and GloVe have revolutionized the way we understand word meanings, enabling the encoding of syntactic and semantic relationships in a way that

computers can process [81, 89]. By transforming words into numerical representations, embeddings facilitate various NLP tasks, including classification, clustering, and similarity measurements. For instance, when analyzing phishing messages, embeddings can help identify similarities between malicious content and legitimate communications, allowing for more accurate detection of phishing attempts.

2.4.2 BERT (*Bidirectional Encoder Representations from Transformers*)

BERT (Bidirectional Encoder Representations from Transformers) represents a significant advancement in the field of NLP by introducing a pre-training approach that enables models to understand the context of words based on their surroundings in a sentence [36]. Unlike previous models that processed text in a unidirectional manner, BERT considers the entire context of a word, leading to a more nuanced understanding of language. This capability is particularly beneficial in identifying deceptive language patterns commonly found in phishing communications.

BERT’s architecture allows it to perform well on a variety of NLP tasks, including text classification, named entity recognition, and question answering. In the context of phishing detection, BERT can be fine-tuned to identify subtle linguistic cues that differentiate phishing attempts from legitimate messages, enhancing the accuracy of detection systems.

2.4.3 *Embeddings in Cognitive Models of Email Decision-Making*

Instance-Based Learning (IBL) models rely on comparing new stimuli with stored past experiences, making the choice of text representation a critical factor in their performance. Early cognitive models of phishing decision-making used semantic similarity measures such as Latent Semantic Analysis (LSA) and WordNet to represent email content [34]. These approaches provided a coarse approximation of semantic relationships but lacked contextual sensitivity, resulting in limited predictive performance (approximately 50–60% accuracy). More recent work has incorporated transformer-based embeddings into IBL frameworks, such as those produced by BERT [144], which improved context awareness but still failed to exceed 60% accuracy. These results suggest that even advanced embedding techniques,

if applied on full email texts, cannot fully capture how humans encode and recall email information.

This performance gap points to a deeper issue: beyond which embedding model is used, there remains the question of *what level of information* should be fed into cognitive models. Prior approaches have predominantly relied on full-text embeddings, implicitly assuming that humans store and retrieve verbatim representations of email content. However, cognitive psychology offers evidence to the contrary. The Verbatim Effect, a well-documented phenomenon in memory research, shows that people are more likely to recall the gist of information rather than its exact wording [103]. For example, after reading a news article, an individual may remember that a storm hit Florida but not the precise phrasing of the report. Similarly, when recalling a supportive comment from a friend, they are more likely to retain the feeling of encouragement than the exact words used.

This suggests that higher-level, abstract representations—such as inferred intent, tone, and other contextual cues—may better approximate how humans encode and recall email communications.

In summary, the exploration of phishing, coupled with an understanding of cognitive processes and memory, lays a crucial foundation for the development of effective defenses against these deceptive practices. By integrating insights from Natural Language Processing, we equip ourselves with powerful tools to analyze and interpret the textual data that underpins phishing communications. The intersection of these fields not only enhances our comprehension of how individuals process information and make decisions in the face of potential threats, but also informs the design of cognitive models that can simulate these processes. As I advance to the subsequent chapters, this interdisciplinary approach will guide my examination of human behavior in phishing contexts and the innovative strategies I can employ to bolster defenses and improve awareness in an increasingly digital landscape.

Chapter 3

SPEARSIM-V2: EXPERIMENTAL DESIGN AND ATTACKER BEHAVIOR ANALYSIS

3.1 Introduction

Phishing attacks, unlike generic spam emails sent for indiscriminate financial gain, are often crafted with varying levels of sophistication and can involve complex, multi-step strategies. The creation of effective phishing emails requires decisions about wording, timing, and context, which suggests that attackers themselves may encounter cognitive limitations similar to those of typical users. Despite this, research on human behavior has largely focused on the challenges that end-users face when recognizing and responding to phishing attempts, with comparatively little attention given to understanding the cognitive aspects of adversarial behavior [37, 38, ?, 58]. Although insights into user difficulties are essential for developing effective security awareness programs, a deeper understanding of the decision-making processes involved in crafting phishing attacks can also inform more robust detection and prevention strategies[1]. Addressing this gap requires tools and environments that enable systematic study of attacker decision-making and behavior in phishing contexts.

Several testbeds have been developed to support experimental research in cybersecurity. For instance, CyberVAN [28] is a high-fidelity virtual network testbed that creates realistic cybersecurity scenarios, making it suitable for testing network security solutions. It has also been used to study how attackers select network nodes to target [4]. However, operating these testbeds often requires dedicated teams to manage services and scenarios. Additionally, there are CyberRanges and security games designed for education and training [128], and comprehensive surveys on cybersecurity testbeds have been conducted, offering taxonomies for classifying CyberRanges based on their scenarios and applications [146]. On the other hand, several testbeds have been created to study how end-users respond to phishing attacks [108, 29]. These simulations mimic email interfaces and expose participants to

a mix of genuine and phishing emails, measuring their responses under varying conditions. However, these environments are not designed to study the behavior of attackers crafting phishing and spear-phishing emails. Similarly, there are some testbeds focused on cyber defenders, such as abstract environments used to study how defense analysts, both individually and in teams, detect cyberattacks and the factors that influence their performance [95, 94]. Yet, these simulations also do not lend themselves to studying the behavior of attackers.

While many studies focus on end-user behavior in phishing scenarios, it is important to investigate the behavior of attackers to understand the strategies they employ while constructing phishing emails. At present, research exploring spear-phishing from an attacker perspective is scarce, limiting our understanding of attackers’ strategies and the cognitive challenges they may face. By gaining insights into their behavior, we can develop more effective defenses against spear-phishing and improve training programs to better address this threat.

3.2 Objective

In the initial phase of my work, I introduce a testbed designed to study attacker behavior in phishing attack scenarios. This testbed is built upon a semi-automated simulation environment based on the SpearSim [143] platform. SpearSim is a multi-player phishing simulation tool originally developed to study both attackers’ and end-users’ behavior in phishing contexts [143, 145]. In this chapter, I will describe the development of SpearSim- V_2 , which builds on the previous work by addressing several challenges and limitations.

Running phishing experiments with human participants can be resource-intensive, requiring substantial time and effort for recruitment, scheduling, and coordination. These challenges also make it difficult to ensure consistency and repeatability across experiments. To address these issues, SpearSim- V_2 replaces human participants with cognitive agents that simulate end-user targets. These agents are designed to model human decision-making processes, enabling controlled and realistic experiments on phishing attacks while reducing the logistical demands of participant management.

Studying attacker behavior at scale is often limited by the availability of human partic-

ipants, making it difficult to run large numbers of trials or explore varied attack scenarios. SpearSim- V_2 addresses this constraint by using cognitive agents in place of human subjects, allowing experiments to be conducted without such bottlenecks. The platform’s scalable design supports testing a wide range of attack and defense configurations in a controlled setting, enabling researchers to examine how skilled social engineers adjust their tactics in response to different security measures and to gain insight into the cognitive processes underlying attacker decisions.

Overall, by developing SpearSim- V_2 I aim to provide a robust, scalable, and controlled platform for studying attacker behavior in phishing attacks, with the ultimate goal of informing the design of more effective defense mechanisms and improving the overall security posture of end-users.

3.3 Method

3.3.1 Building Blocks

The main idea proposed in this chapter is to substitute human end-users with data-driven cognitive models, often referred to as cognitive twins, in studies of spear-phishing attacks. These cognitive twins are developed using well-established cognitive architectures such as ACT-R and Instance-Based Learning Theory (IBLT), combined with natural language processing algorithms that can simulate real-time, ”human-like” responses to phishing attempts.

By integrating advanced cognitive models, researchers can observe how attackers interact with these virtual targets in a more controlled and scalable manner. This method builds on prior work in phishing simulations, such as SpearSim [143], as well as cognitive models of phishing responses based on instance-based learning theory [48]. The result is a more efficient and realistic framework for studying attacker strategies and behaviors in spear-phishing contexts.

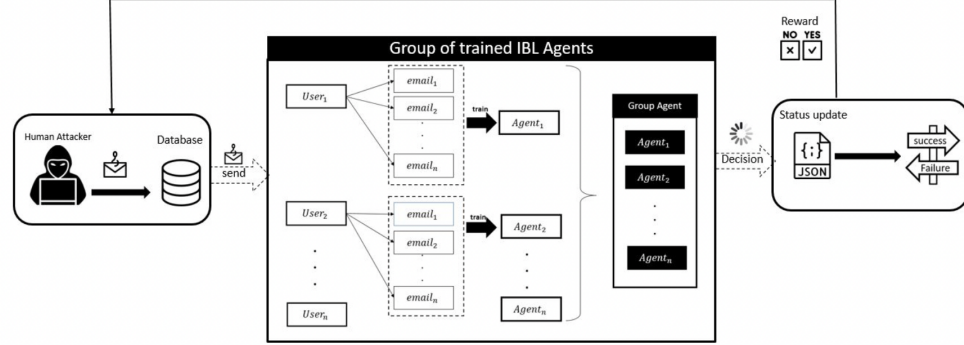


Figure 3.1: Integrating Software and the Cognitive Model Procedure Overview

The SpearSim- V_2 simulation is composed of four key components: 1) an Attacker Interface to simulate the tasks and actions involved in generating phishing attacks, 2) a Natural Language Processing Module to represent phishing emails in high-dimensional vectors, 3) Cognitive Agents to predict human responses to spear-phishing emails, and 4) Software modules to integrate the attacker interface with the cognitive agents (Figure SpearSim serves as a human-in-the-loop simulation that recreates spear-phishing tasks, with cognitive models (IBL Agents) acting as end-user agents. I then discuss how these cognitive agents were integrated into SpearSim, detailing the technical processes and software elements used to develop the semi-automated simulation environment.

3.3.2 Phishing Simulation:

SpearSim is a computer-based simulation tool designed to replicate the actions and cognitive processes involved in executing spear-phishing attacks, a more targeted form of phishing. In the previous version of SpearSim, conducting lab experiments required four human participants, where one acted as the attacker and the other three as end-users. The attacker would create phishing attempts aimed at the end-users without their knowledge. SpearSim was specifically developed for studying attacker behaviors in spear-phishing scenarios. For a detailed description of the software, refer to the original paper [143].

3.1).

Phishing Attack Console

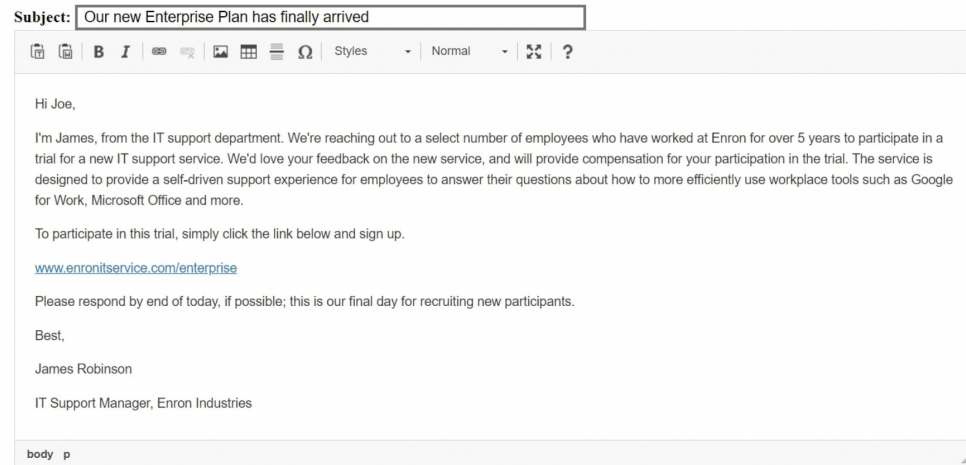


Figure 3.2: Attacker Interface with an example attack

We are waiting for a response from your target



Please wait for the response from your target... it may take a few minutes...

Do not refresh!

Trial:2

Remember: In a Phishing attack, there are no certainties on rewards from each attack. Be persistent to score big

Figure 3.3: Attacker interface after attack deployment



Figure 3.4: Possible Responses provided by the software

Conducting experiments with SpearSim faced several challenges and limitations. A significant issue was the need to recruit and schedule four participants for each experiment, including one skilled in red-teaming and social engineering, and three others for the end-user roles. This recruitment difficulty often led to using non-expert participants. Automating the end-user roles with cognitive agents provides flexibility, allowing studies to focus on skilled attackers. Additionally, using human participants as targets meant excluding malicious payloads due to safety concerns, a limitation resolved by cognitive twins. Lastly, while the previous SpearSim version limited attackers to three end-user targets for convenience, cognitive twins remove this constraint, allowing attackers to engage with multiple agents and profiles in each experiment.

3.3.3 Cognitive Agent:

SpearCog [144] is a cognitive agent designed to predict human responses to email messages, particularly in the context of phishing and spear-phishing attacks. The model is based on Instance-Based Learning Theory (IBLT) [48] and was developed using Python (PyIBL, [84]). The primary goal of SpearCog was to explore how cognitive processes, specifically memory activation and retrieval, influence end-user reactions to phishing attempts. While it was initially developed for theory-building purposes—examining the cognitive mechanisms that contribute to phishing susceptibility—it is also a computational model capable of simulating

human responses and can be used for real-time phishing prediction. For a full description, please refer to the original paper [144].

Natural language models are essential for representing email instances within cognitive agents. The Bi-directional Encoder Representations from Transformers (BERT, [36]) model is employed as the primary approach to encode email text into high-dimensional vectors. BERT’s bi-directional nature allows it to consider the full context of a word by analyzing both the preceding and succeeding words simultaneously. This is achieved through its use of a transformer-based self-attention mechanism, as proposed by [131]. To integrate email representation into the cognitive agents, the Sentence-BERT (SBERT) architecture was utilized, specifically employing a distil RoBERTa model fine-tuned on the SNLI corpus [21]. The semantic similarity between email pairs was calculated using cosine similarity, a crucial step for modeling memory activation in Instance-Based Learning (IBL) models based on similarity.

I integrated SpearSim [143] with cognitive agents, or cognitive twins, designed to model and predict individual human responses to phishing attacks. Unlike conventional machine learning models, which typically analyze population-level data, cognitive agents can account for individual variability in phishing susceptibility. This allows the model to capture personalized differences, where some individuals are more vulnerable to certain types of spear-phishing attacks, while others may remain unaffected.

Figure 3.1 illustrates the architecture of the SpearSim- V_2 simulation system. In this setup, a skilled attacker participant, such as a professional red-teamer or social engineer, can design and launch spear-phishing emails targeting cognitive agents. These agents represent human profiles created specifically for the experiment, using information about human behavior and responses drawn from previous studies [143]. Once the attacker crafts the spear-phishing email, it is stored in a relational database (e.g., MySQL) with a predefined schema. Based on the agent profile selected by the attacker, the email is forwarded to the corresponding cognitive agent, which simulates a "human-like" response to the phishing attempt (see the central section of Figure 3.1).

The cognitive agent predicts whether the response will be to "respond" or "ignore" the email. This prediction is stored in a .json file that logs the status of each attack trial during

the experiment. If the cognitive agent "responds," it represents a successful phishing attack, while an "ignore" prediction represents a failed attack. The attacker interface then provides feedback, indicating the success or failure of the attack, and updates the attacker's reward accordingly. Figures 3.2 to 3.4 depict the spear-phishing console, the simulation of the attack, and the automatic responses generated by the system.

3.3.4 Procedure

All computations were performed on a system featuring an Intel Core i7-7500U CPU, using Python 3.7.15. The attacker interface and simulation software were hosted on a local server environment, running MAMP with PHP 7.1.23. The cognitive agents were developed by training IBL agents on a dataset of 529 unique emails from a previous experiment. This dataset included responses from 84 participants and consisted of 210 ham emails, 15 mass phishing emails, 52 promotional emails, and 252 spear-phishing emails. Cognitive agents were created to represent the behavior of each participant, enabling near real-time response predictions within the simulation.

The participants' original responses were encoded as one of five actions: "respond immediately," "flag for follow-up," "leave in inbox," "delete," or "delete and block sender." For each email, the cognitive agents retrieved past similar instances to generate a simplified response of either "respond" or "ignore." A successful attack was indicated by the first two responses, while the remaining three represented failed attacks. The cognitive agents were trained using an 80-20 train-test split ratio for response prediction.

3.4 Results

In this section, I present the evaluation results of SpearSim- V_2 . In this study, I replicated the previous experiment, but this time measured the responses of cognitive agents in place of human end-users. The two versions were compared on the basis of two key metrics: accuracy and response time.

Accuracy: In this replication, cognitive agents were used to predict the responses of human participants from the previous experiment. Accuracy was evaluated by determining the percentage of emails for which the agents' decisions matched those of the humans. The

cognitive agents achieved an accuracy of 57.9%, aligned with human responses for more than half of the emails.

Response Time: For response time, I measured the duration from the moment the attacker sent the message until the cognitive agents produced their response. This included the time required by the NLP module to generate the vector representation of the email and the time taken by the cognitive agent to make a decision based on previous experiences. On average, the cognitive agents responded in 16.9 seconds (standard deviation: 0.38), compared to the human participants, who had an average response time of 45.6 seconds (standard deviation: 55.2).

A paired t-test was conducted to compare the response times of the cognitive agents and human participants, showing a significant difference with a p-value of less than 0.001. Figure 3.5 illustrates the differences in response time distributions between the two groups. Human participants exhibited greater variability in response times, following an exponential distribution with a long tail, indicating a higher average, greater variance, and more outliers. In contrast, the response times of cognitive agents followed a more uniform distribution, reflecting more consistent responses with lower variability, centered closely around the median.

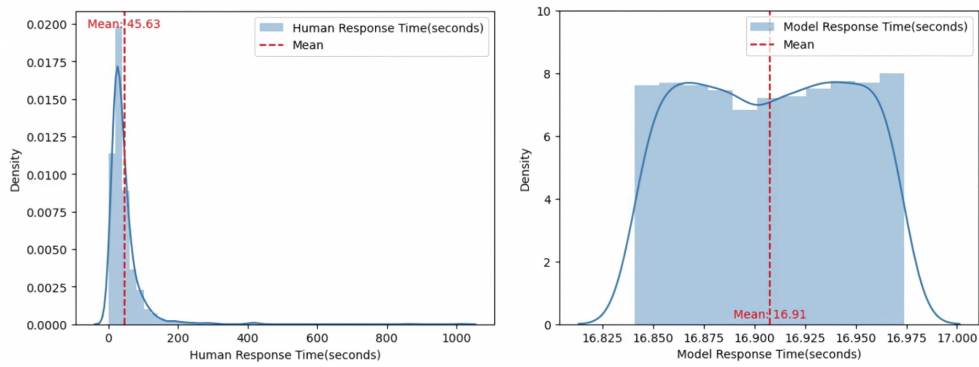


Figure 3.5: Software and Human Response Time

3.5 Discussion

This chapter presented SpearSim- V_2 , a synthetic task environment developed for conducting experiments on attacker behaviors in the context of spear-phishing attacks. I evaluated the performance of SpearSim- V_2 against its predecessor using two key metrics: accuracy and response time. In terms of accuracy, SpearSim- V_2 delivered encouraging results. The software responses were approximately 60% aligned with the human responses, and certain agents outperformed others to accurately predict the behavior of human participants. These findings indicate that cognitive agents have the potential to effectively replace humans in end-user roles. However, some agents were less successful in predicting human responses, suggesting that future versions of the simulation may require alternative approaches to better model human behavior. As for response time, SpearSim- V_2 significantly outperformed its predecessor. By substituting human users with cognitive agents, I reduced the average response time to 16.9 seconds, even with standard computing resources, whereas human participants averaged 45.6 seconds with high variability. This variability can greatly influence experiment logistics, as longer wait times could frustrate participants in the attacker role, potentially leading to higher dropout rates.

The earlier version of SpearSim was a four-player simulation where three participants took on the role of end-users. In contrast, my updated version offers increased flexibility, allowing us to incorporate a wider range of agents as end-user targets. This enhancement enables experiments that are more ecologically valid and scalable. Importantly, this implementation made it possible to evaluate attacker strategies and model user responses without requiring human participants for every experiment, a major step forward to conducting controlled repeatable phishing studies in the future.

For this study, I opted to use cognitive models over random models for several reasons. Although the specific model used may not have been crucial for this initial evaluation, it is highly significant for future research on this platform. In particular, for upcoming studies involving skilled social engineers, the cognitive model allows for the creation of more “human-like” responses. Ensuring that attackers perceive these responses as realistic, rather than random, enhances the experiment’s credibility and makes it a more accurate

reflection of real-world spear-phishing scenarios. This realism may also influence the tactics employed by the attackers during the experiment.

3.6 Limitations and Future Work

While the development of SpearSim- V_2 represents a substantial step forward, some limitations emerged that informed the direction of the next phase of my research. First, my evaluation focused only on the accuracy and response time metrics; Expanding future assessments to include the precision, recall, and F1 scores would provide a more comprehensive understanding of the software performance.

Second, while the simulation captured basic response patterns, the IBL model showed limited performance and did not exceed an accuracy of 60%, indicating that it often missed key decisions humans would make. This gap suggests that current IBL models relying on full email texts do not fully reflect how humans recall and interpret phishing emails. The models used in this simulation, rely on memorizing entire texts rather than approximating the more selective and abstract nature of human memory. This limitation highlights a critical need: cognitive models must use representations that better align with the way people encode and recall email content when making decisions.

These observations guided the next phase of my research: developing alternative representations of phishing emails using high-level, human-annotated cues derived from existing survey data. This direction aims to create representations that not only improve model performance but also more closely align with human recall and decision-making processes. This insight became the motivation for the next chapter, which introduces a privileged learning framework to leverage human-provided cues for improved cognitive modeling.

Chapter 4

PRIVILEGED LEARNING FOR MODELING HUMAN DECISION-MAKING IN PHISHING ATTACKS

4.1 Introduction

The previous chapter demonstrated that Instance-Based Learning (IBL) models can approximate human responses to phishing and non-phishing emails when using raw email embeddings as input representations. However, these models achieved only moderate alignment with human decision-making, suggesting a critical gap: representing emails as high-dimensional text embeddings does not fully capture how humans actually recall and evaluate email content.

Cognitive psychology offers insight into this limitation. People rarely retain verbatim details of long, complex texts; instead, they encode and recall the *gist*—the core meaning, intent, or topic of messages, abstracting away surface-level wordings [103]. This phenomenon implies that cognitive models may benefit from using human-interpretable higher-level features that reflect what people consciously remember about an email conversation rather than the full raw text of the emails.

This chapter represents a first step toward converting raw email text into lower-dimensional, cognitively meaningful representations. To achieve this, I use existing survey-derived features that capture the perceptions of participants about email topics. These features provide a condensed representation of emails aligned with how individuals consciously interpret their content. By integrating these human-informed features into IBL models, this chapter examines whether such representations improve the models’ ability to simulate human decision-making when evaluating phishing and non-phishing emails.

Specifically, this chapter addresses two key research questions:

1. Do higher-level, survey-derived cues improve the performance of IBL models compared to using raw text embeddings?

2. Can we generate these survey-style features for new emails where direct human annotations are unavailable?

Through these investigations, this chapter aims to evaluate whether incorporating human-informed, cognitively aligned representations can improve predictive accuracy and better reflect the mental processes underlying email-based decision-making.

4.2 Objective

Building on the limitations identified in the previous chapter, the objective of this chapter is to investigate whether cognitive agents that use user reported survey-derived features, reflecting how individuals interpret the objectives and topics of emails, can better predict human responses to phishing and non-phishing emails. By incorporating privileged information from participant surveys into models based on Instance-Based Learning Theory (IBLT), this work explores whether discrete, human-informed representations can improve the alignment between model predictions and actual human decision-making. Through this approach, the chapter seeks to evaluate the potential of using features that capture what people consciously recall about emails as an alternative to modeling decisions based solely on raw email text.

4.3 Method

This section describes the methodology used in this study, providing an overview of the core components of the modeling framework and a clear explanation of the data preparation process.

4.3.1 Dataset and Experiment

The dataset analyzed in this study originates from a previous laboratory experiment designed to examine how end-users process and respond to a variety of emails in a simulated decision-making environment [144]. In this experiment, participants were organized into groups of four, with three participants acting as end-users tasked with evaluating emails and a fourth participant assigned a different role within the scenario.

Each end-user was provided with a detailed fictional narrative and instructed to make decisions on behalf of their assigned character regarding a series of emails. These emails spanned multiple categories, including legitimate, promotional, mass phishing, and targeted phishing messages, to offer a diverse range of decision-making contexts for analysis.

The dataset comprises 395 emails, each reviewed by a varying number of participants, with up to 59 raters evaluating a single email and an average of 18 raters per email. Participants responded to a series of questions about each email, such as whether they would trust it or choose to respond. In addition, they completed a survey that assessed their perception of the topic of each email, covering factors such as whether it requested an action, referenced a deadline, was perceived as spam or related to a meeting, among other attributes. These survey responses provide privileged information on what participants consciously noticed and recalled, offering valuable insights for developing cognitive models that more accurately simulate human decision-making processes when evaluating emails.

This model integrates cognitive modeling techniques with privileged learning [130], enabling it to incorporate higher-order, lower-dimensional representations of email content based on human perceptions. Traditional approaches often fed the full email text into cognitive agents to calculate similarity using measures such as cosine similarity [144], but these methods may not align with how humans selectively recall information when processing emails. In contrast, the approach in this chapter uses discrete representations derived from survey responses collected from participants, which capture the aspects of emails that individuals consciously noticed and remembered. By modeling these human-informed features within an IBL framework, this method aims to better simulate how people recall and respond to emails during decision-making.

4.3.2 Feature Extraction and Label Aggregation

In this stage, I aggregated the survey responses for each email applying majority voting across participants. For each email, a binary matrix is created, where rows correspond to individual emails and columns represent responses to specific survey items. This matrix is constructed by comparing the aggregated responses with the total number of raters per

email, resulting in a set of privileged characteristics that capture the most commonly perceived attributes of each message. This aggregation addresses the multi-label nature of the dataset, where each email may have multiple labels assigned by different participants. By consolidating responses into a single feature vector per email - reflecting the dominant perceptions reported by participants - and determining a consensus label vector through majority voting, the data set is transformed into a format suitable for modeling within the instance-based learning framework.

The model developed in this chapter addresses the dataset’s multi-label nature, where each email can be associated with multiple human-assigned labels reflecting different perceived attributes. By incorporating privileged features derived from aggregated survey responses, the model represents each email with a vector that captures these key perceptions. The instance-based learning framework then uses these representations to simulate how humans might recall and make decisions about emails, allowing the model to evaluate its ability to predict participant responses across multiple categories of labels.

4.3.3 *Model Architecture*

The predictive model is a sequential neural network built using the Keras library[30], designed to estimate human survey responses directly from email content. This network serves to generate privileged features reflecting what humans consciously notice in emails, which are later used in the instance-based learning framework.

The architecture consists of three main layers:

- **Input Layer:** Contains 64 neurons with a ReLU activation function, with the input dimension equal to the length of the email embeddings. The ReLU activation introduces non-linearity, enabling the model to capture complex patterns in the email text.
- **Hidden Layer:** Comprises 32 neurons with a ReLU activation function, allowing the network to learn deeper, more abstract representations of the input.

- **Output Layer:** Includes one neuron for each survey question (column) with a sigmoid activation function for multi-label binary classification, outputting probabilities representing the likelihood of each perceived attribute being present in the email.

4.3.4 *Neural Network Model Training and Evaluation*

The neural network model is trained on the training set for up to 100 epochs using a batch size of 32. To avoid overfitting and improve the model’s ability to generalize to unseen data, an early stopping mechanism is implemented. This mechanism monitors the validation loss during training and stops the process if no improvement in the validation loss is observed for 10 consecutive epochs. If the validation loss increases, indicating a potential overfit, the training is stopped and the model weights are restored to those of the epoch with the lowest validation loss. After training on the privileged survey features, the model is applied to the test set to generate probability estimates for each feature, reflecting the likelihood that an email possesses that feature. These estimates represent the model’s inferred survey responses for use in further analyses. This strategy ensures a well-calibrated model that generates features closely aligned with human perceptions.

4.3.5 *Aggregated Agent Decisions*

In this study, one IBL agent is trained for each individual human rater in the dataset, resulting in a total of 84 agents. Each agent learns from the emails and corresponding responses of a single participant, capturing the unique decision-making patterns of that rater. When a new email is presented during testing, each agent independently predicts a response based on its learned experiences. The final predicted response for the email is then determined by a majority vote across all agents’ outputs. This approach enables the model to simulate the diversity of human perspectives and aggregate them into a collective prediction reflecting broader human decision-making behavior.

The decision to train an individual agent for each human rater was made to capture the unique decision-making patterns of all participants in the dataset. This resulted in 84 agents, one per rater. While some raters provided extensive responses across many

emails, others evaluated only a small number of emails, which could limit the predictive power of their respective agents. Nevertheless, including all raters allows the modeling framework to account for the full diversity of human perspectives present in the dataset. By aggregating predictions from all agents through majority voting, the system can balance individual variability and produce collective predictions that better reflect overall human decision-making behavior.

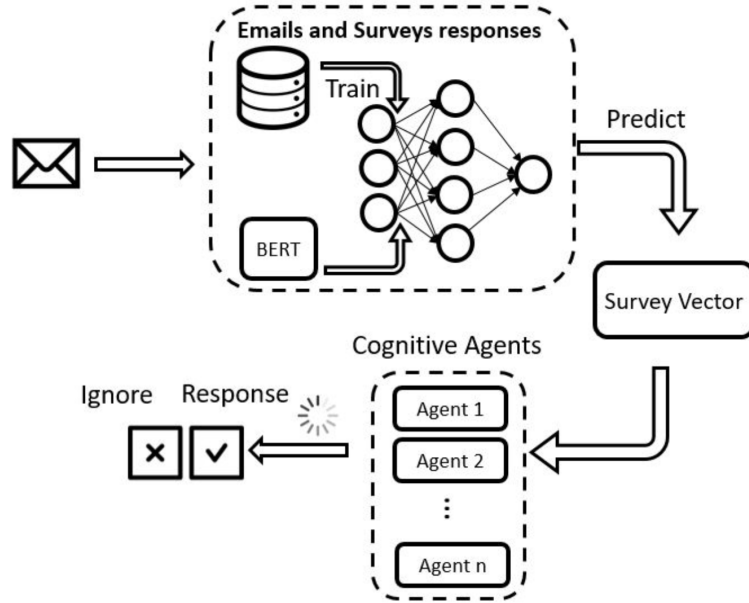


Figure 4.1: Procedure

The implementation of each IBL agent in this chapter follows the standard activation-based retrieval mechanisms established in Gonzalez et al. [48], detailed activation formulas and modeling parameters are presented in chapter 2, in sub-section 2.2.2.

In summary, this approach combines the strengths of neural networks and instance-based learning to develop cognitive agents that model human responses to emails. A neural network predicts survey-derived feature vectors for new emails, simulating the key aspects humans consciously notice and recall. These predicted features are then used by instance-based learning agents, each trained on individual raters' data, to simulate personalized

decision-making patterns. By aggregating the predictions of all agents through majority voting, the model produces collective responses that closely approximate human judgment. This methodology provides a cognitively grounded framework for understanding and predicting how people interpret and respond to phishing and non-phishing emails. Figure 4.1 illustrates the overall procedure for building the proposed framework.

4.3.6 Prediction and Evaluation

The performance of the cognitive agents is evaluated using two distinct datasets. The primary dataset, which has been the focus of this discussion, is divided into an 80-20 train-test split: 80% of the data is used for training the agents, while the remaining 20% is reserved for testing their ability to predict participant-like responses to emails. This evaluation framework allows for assessing how effectively the agents, trained on survey-derived feature representations, can simulate human decision-making patterns when presented with new email instances. Model performance is quantified using standard classification metrics, including accuracy, precision, recall, and F1-score, to evaluate how closely the predicted responses align with actual participant judgments.

In addition to the original dataset, I introduce a second dataset consisting entirely of new, unseen emails collected from a prior laboratory study [117]. This dataset differs from the first in format and topical content and comes from a different source. While the cognitive agents were trained exclusively on the original dataset, they were tested on this new dataset to evaluate how well the modeling approach generalizes to emails outside the training domain. This second dataset includes responses from 48 participants, totaling 3,840 responses to 241 unique emails.

On average, each participant responded to approximately 80 emails. The dataset includes both benign emails (commonly referred to as “ham”) and phishing emails, with 55 labeled as benign and the remaining 186 labeled as phishing. Evaluating the agents on this new dataset enables assessment of their ability to generalize predictions of participant-like responses to previously unseen emails. Model performance is measured using standard classification metrics, including accuracy, precision, recall, and F1-score, to quantify how

closely the agents’ predictions align with actual participant judgments.

4.4 Results

In this section, I present the findings of my research, highlighting the predictive performance of the Instance-Based Learning (IBL) agents trained on privileged, survey-derived features. These results demonstrate the agents’ ability to approximate human decision-making patterns when evaluating phishing and non-phishing emails. The insights gained provide a deeper understanding of how well different modeling approaches capture participant-like responses, offering evidence of the alignment between model predictions and actual human judgments.

Neural Network Performance:

In the initial phase of this study, I evaluated the performance of the neural network component, which is designed to predict feature vectors derived from surveys for emails. These predicted vectors serve as privileged features for subsequent modeling within the instance-based learning framework. Model performance was assessed across multiple epochs by recording the training and validation loss, as well as the accuracy for each epoch.

I implemented an early stopping mechanism to mitigate the risk of overfitting. This mechanism monitors the validation loss during training and halts the process if there is no improvement in the validation loss for a specified number of epochs, which in this case was set to 10. Consequently, the training was concluded at the 40th epoch due to this early stopping criterion.

Throughout these 40 epochs, the model demonstrated a steady decline in validation loss, indicating effective learning from the training data. Accuracy improved consistently over time, reaching a final training accuracy of 86.97% and a validation accuracy of 87.09%. These results show that the neural network was able to accurately predict survey-derived feature vectors for emails in both the training and validation sets. The close alignment of training and validation accuracies suggests that the model generalizes well to unseen data, an essential attribute for reliably simulating participant-like responses in subsequent analyses.

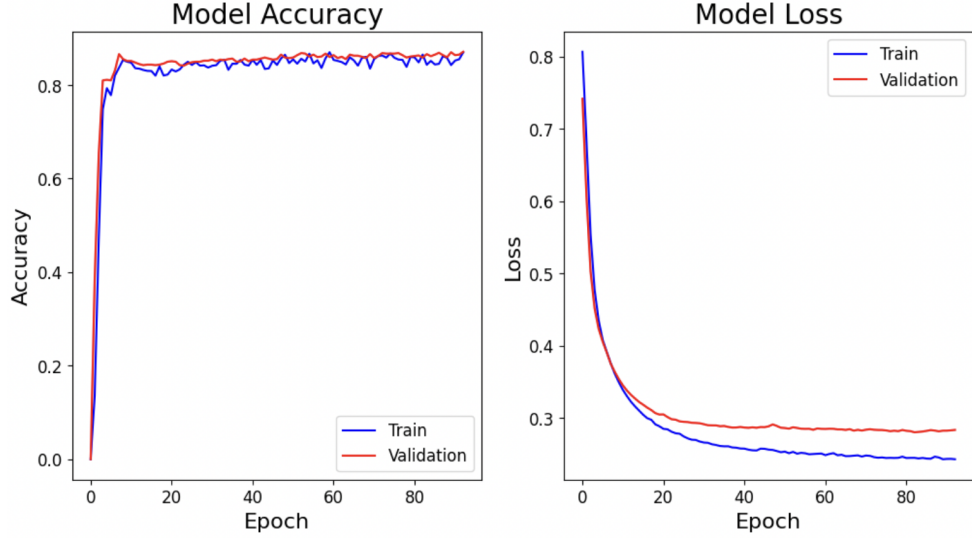


Figure 4.2: Model Accuracy and Loss for training and validation

Overall, the neural network demonstrated strong performance in predicting survey-derived feature vectors from emails. The consistent reduction in validation loss and high final accuracy highlight the model’s effectiveness in transforming high-dimensional email content into lower-dimensional representations aligned with human perceptions. This transformation is a crucial component of the methodology, enabling subsequent cognitive modeling analyses to operate on features that more closely reflect what participants consciously recall when evaluating emails. Additionally, the early stopping mechanism contributed to maintaining both accuracy and robustness, supporting the model’s ability to generalize to new, unseen data.

4.4.1 Cognitive Agent Performance

Building on the neural network’s predicted survey features, the next phase of the approach involves incorporating IBL agents. Each agent is trained on a distinct subset of data corresponding to a single participant, allowing it to model that individual’s unique decision-making patterns. This individualized training process enables each agent to develop a fine-tuned representation of how a specific participant evaluated emails, enhancing the model’s

ability to capture the diversity of user behaviors and perspectives when responding to both phishing and benign emails.

The decision-making process for these agents is guided by two primary elements. First, each agent relies on instances stored in its memory, representing the participant’s previously observed responses to emails. Second, the agents use the survey-derived feature vectors predicted by the neural network for new emails. These vectors serve as condensed yet meaningful representations of the emails, enabling agents to generate participant-like responses without needing to process the full complexity of the raw email content. This streamlined approach allows each agent to simulate individual decision-making patterns based on both past experience and new input.

Importantly, these strong performance metrics were achieved without directly incorporating the full text of the emails into the cognitive models. Instead, lower-dimensional survey-derived feature vectors were used as alternative representations. This approach demonstrates the potential of compact representations to capture critical information relevant to human decision-making processes. By transforming detailed email content into concise vectors aligned with participant perceptions, the model is able to simulate human responses more effectively while reducing complexity and processing demands.

The cognitive agents’ performance is evaluated using two distinct datasets to comprehensively assess their ability to simulate participant-like responses. The primary dataset, which was used for training, is split 80-20 into train and test sets, enabling evaluation of the agents’ effectiveness in predicting responses to familiar emails. The second dataset, consisting entirely of new, unseen emails that were not part of the training phase, adds complexity to the evaluation. This setup allows for assessing the agents’ capacity to generalize their learned decision-making patterns to novel email instances, highlighting their ability to model human responses in previously unencountered scenarios.

The agents’ performance was evaluated using accuracy, precision, recall, and F1 score metrics. On the first original dataset, the agents achieved an accuracy of 73%, a precision of 63.1%, a recall of 67%, and an F1 score of 65% on the testset. On the second dataset, which contained entirely new emails from a different source, the agents achieved an accuracy of 62%, a precision of 58%, a recall of 60%, and an F1 score of 59%. These results demon-

strate that the cognitive agents, trained on individualized participant data, are capable of approximating human decision-making patterns with reasonable accuracy, even when applied to novel emails outside of their training experience. This supports the effectiveness of the proposed approach in modeling human responses to both phishing and benign emails.

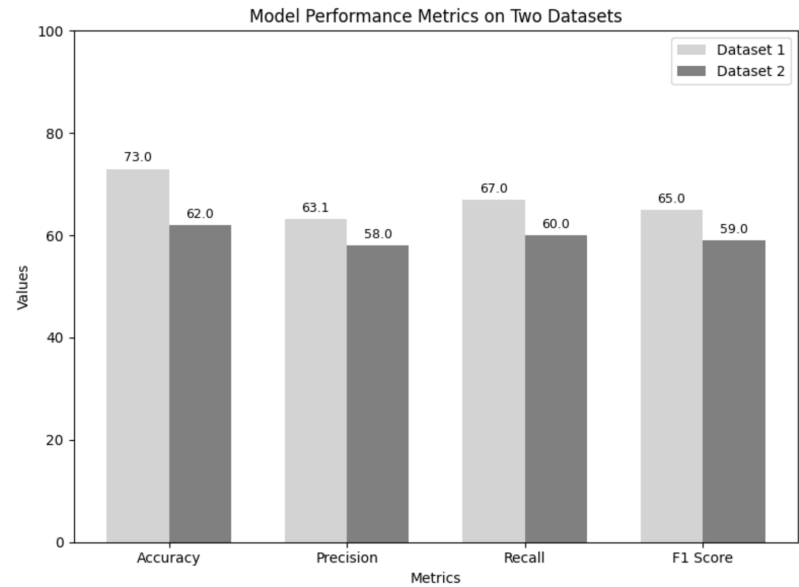


Figure 4.3: Model Performance Metrics on Two Datasets

In summary, these findings highlight the effectiveness of the proposed approach in modeling participant-like responses to emails. By combining the predictive capabilities of a neural network with the individualized decision-making patterns of cognitive agents, this methodology captures key aspects of how people interpret and respond to both phishing and benign emails. While the model demonstrates reliable performance on familiar data, it shows a noticeable drop in accuracy when applied to previously unseen emails, indicating challenges with generalizability that warrant further investigation. These results nonetheless support the potential of this cognitively grounded framework for advancing our understanding of human decision-making in email contexts.

4.5 Discussion

This study introduces a new approach to understanding how individuals assess emails by modeling the relationship between new email stimuli and their prior experiences. By combining cognitive modeling with machine learning, this work offers a novel perspective on simulating human decision-making when evaluating phishing and benign emails. A key contribution of this research is embedding high-level, lower-dimensional contextual representations—derived from human survey responses—into cognitive models, enabling more manageable yet informative feature vectors. This strategy enhances cognitive agents’ ability to approximate how humans recall and interpret emails, bridging the gap between cognitive modeling and machine learning representations. The integration presented in this study provides an innovative method for representing and processing lengthy texts within cognitive architectures, advancing our understanding of human responses to emails in diverse contexts.

4.5.1 Key Benefits

This study demonstrated two notable benefits of integrating survey-derived features into cognitive models:

- **Neural networks can generate survey-style responses for new emails with surprisingly high quality.** The model achieved strong performance in predicting participant survey responses for previously unseen emails, showing that machine learning can effectively extend survey-based annotations to new data.
- **IBL models perform better when using these high-level survey features instead of raw full-text embeddings.** By replacing full-text inputs with structured, survey-derived vectors, the cognitive agents more closely approximated participant-like decision-making.

These findings suggest that alternative text representations can provide cognitively meaningful inputs for modeling human decision-making processes, while also offering practical advantages in handling high-dimensional email content.

4.5.2 *Key Limitations*

Despite these promising results, the study revealed two critical limitations:

- **Label specificity.** The survey-derived features were created from a fixed set of labels tied to a single dataset. These labels primarily captured participants' perceptions of email topics but may not fully reflect the range of cues people actually use when recalling and evaluating diverse emails.
- **Poor generalization.** When tested on a different dataset, the model's performance declined. This drop highlights the challenge of applying narrowly defined survey features to new content, limiting the approach's ability to generalize beyond the original dataset.

Additionally, the survey responses were encoded using one-hot vectors, which fail to capture relationships between features and do not align with the gist-based, interconnected nature of human memory. These representational constraints likely contributed to the observed performance drop when working with new email content.

4.5.3 *Implications and Future Directions*

These findings highlight the need for a more systematic and cognitively grounded approach to email representation. One that:

- Is more generic and is not specific to a single dataset,
- Better reflects how humans encode, recall, and evaluate email content, and
- Captures relationships between features to approximate the interconnected structure of human memory.

To overcome these limitations, the next chapter systematically explores a variety of alternative text representations ranging from lower-level features, such as summaries and keywords, to higher-level abstract features, such as inferred intent, sentiment, and tone.

These representations are designed to match more closely the way people process and recall messages across different topics, ultimately improving the alignment between cognitive models and human decision making.

Chapter 5

INSTANCE REPRESENTATIONS THROUGH INTENT MODELING

5.1 Representation Strategies for Modeling Human Recall

Building on insights from the previous chapter, where designing instance representations with privileged information revealed the importance of aligning instance features with what humans would encode and recall to memory, this chapter systematically investigates intent modeling as a methodology for representing email text as instances within IBL models. Previous experiments with IBL models of phishing decision making indicated that long-form email representations such as full-text embeddings often degraded the model’s ability to approximate participant-like responses [144]. Although these embeddings capture detailed semantic information used in conversation [89], they do not reflect how humans process and remember conversations [79]; *Individuals tend to retain only a few key ideas and impressions rather than entire messages* [82, 64, 9, 78, 22].

This chapter builds on this key idea and addresses the challenge of instance representation by analyzing the effectiveness of email intent modeling as an approach for instance representation [139]. In this chapter, I systematically compare two broad types of representations containing:

- Lower-level text-based features, including full-text embeddings, embeddings for email summaries, and embeddings for keywords.
- Higher-level abstract features, such as perceived intent, sentiment, and tone.

By evaluating how IBL models with each type of representation align with human response decisions to emails including phishing and non-phishing emails, this work aims to advance our understanding of how to represent text-based decision stimuli as instance representations in cognitive models, enabling more accurate simulations of human recall processes when evaluating susceptibilities to phishing threats.

Previous works built IBL models with emails represented as a single attribute-value pair, typically using the full message or sentence embedding as the primary feature representing each email. This “flat” representation limits the model’s ability to model higher-order similarity patterns that people would use when retrieving related conversations, such as considering two emails that request a password reset as similar in intent despite using different words or phrasing. In my previous study, I found that using the responses of the self-reported surveys of the participants, specifically their perceptions of the topic of each email, as representations of the emails was effective, highlighting the value of incorporating features human-interpretable of higher order [79, 144, 34].

Cognitive psychology suggests that human memory does not encode and retrieve text verbatim but instead relies on hierarchical recall mechanisms based on the gist of information [82, 64, 9, 78, 22]. Through processes such as chunking and abstraction, individuals tend to remember salient elements, such as emotional tone, communicative intent, or key phrases, while disregarding less relevant details present in text. Based on these findings, I hypothesize that, rather than recalling the entire text of an email, people are more likely to retain overarching intentions or a few critical keywords that capture its essence. Consequently, using human-interpretable higher-level features as instance representations should improve the ability of IBL models to simulate how people recall and respond to emails.

5.2 Intent Modeling

Intent modeling is the process of computationally identifying the underlying communicative purpose or objective behind a message, such as making a request, offering information, or expressing urgency [26, 149]. In natural language processing (NLP), intent analysis enables systems to interpret user goals and provide contextual responses, which is fundamental to optimize dialogue systems, search engines, and recommendation systems [149, 42]. This capability is also critical in applications such as email assistants, where understanding intent improves message prioritization, task automation, and user experience [86, 42].

Intent modeling in the email domain has evolved through both linguistically grounded and machine learning-based approaches[31, 106, 72]. Early work applied speech act theory to categorize communicative functions such as request, commit, or deliver within emails,

demonstrating the feasibility of intent detection based on message-level annotations rather than sentence-level labeling [31, 106]. Building on this foundation, later studies leveraged neural network architectures to model actionable email intents such as scheduling or task management [72]. Others have incorporated contextual information from email threads to improve intent classification accuracy, showing that message context significantly improves model performance [139]. More recently, weak supervision has enabled the training of scalable intent classifiers by generating labels from patterns and heuristics rather than manual annotation [115]. In this setting, patterns and heuristics refer to rule-based labeling functions derived from user interactions—for example, heuristically labeling any email that receives a reply with an attachment as a request intent—providing noisy but scalable supervision signals that complement a small manually annotated dataset” [115]. These diverse methods underscore the importance of modeling communicative intent in emails to support prioritization, task automation, and cognitive offloading for users [72].

In this study, I move beyond using intent modeling solely for assigning classification labels and instead employ the learned intent representations themselves. These representations encode higher-order communicative features that serve as inputs to cognitive models of human decision-making.

This chapter evaluates intent modeling as an approach for instance representation for cognitive models of phishing decision-making supported by prior research demonstrating that communicative intent influences perceptions of message credibility, urgency, and susceptibility to phishing [24, 141, 109].

I hypothesize that representing emails using email intents as instances leads to IBL models that more effectively align with human response decisions, particularly in the context of phishing and spear phishing emails.

To test this hypothesis, I compare the following two categories of representational approaches within the IBL framework:

- **Lower-Level Text-Based Representations:**

- **IBL + Full-Text:** This is the same approach used in previous studies [144] and serves as the baseline in this chapter for evaluating whether more cognitively

aligned representations can improve model performance.

- **IBL + Summaries:** Using large language model (LLM)–generated summaries as a more compact alternative to full email texts, providing a reduced representation while retaining the main informational content.
- **IBL + Keywords:** Using self-reported keywords that participants identified as the terms they would use to recall each email in their inbox, reflecting a gist-based representation of the message.

- **Higher-Level Abstract Representations:**

- **IBL + Intents:** Using human-rated communicative intent labels to represent each email, capturing the overarching purpose or action implied by the message (e.g., requesting information, confirming an event). This approach aligns with gist-based memory by abstracting away detailed wording in favor of the core intent perceived by recipients.
- **IBL + Intent, Sentiment, and Tone:** Using a combination of human-rated communicative intent, emotional tone, and sentiment labels to create a richer, higher-level representation of each email. This multidimensional abstraction captures not only the perceived purpose of the message but also its emotional and interpersonal cues, providing a more comprehensive gist-based representation aligned with how people evaluate and recall emails.

All of these representations—whether text-based or abstract—are converted into embeddings before being stored and compared within the IBL framework. Figure 5.1 illustrates the progression of these representation types, ordered from lower-level text-based inputs to higher-level abstractions

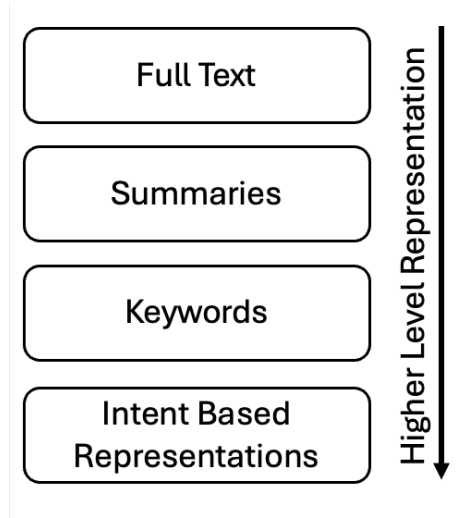


Figure 5.1: Hierarchy of email representations ranging from lower-level to higher-level abstract representations

The remainder of this chapter describes the data collection process, modeling setup, evaluation metrics, and comparative results between the evaluated representation strategies.

5.3 Data Collection:

I conducted a study with human subjects on the Prolific platform to collect features such as email summary, search keywords, tone, sentiment, and intent that participants reported being the most salient during recall and email response. In the study, participants received a series of emails from previous phishing experiments [143] and were asked to evaluate summaries of emails produced using generative AI, self-report keywords they would use to recall each email, and annotate emails with perceived tone, sentiment, and intent. The data collected from this study was used as features, dimensions along which instance similarity was computed in the IBL model, as described in Sections Methods (Section 5.4) and Evaluation (Section 5.5).

5.3.1 *Email Dataset:*

The emails the participants received included a mix of primary messages, promotional content, phishing, and spear phishing emails. These messages were drawn from a previous study described in [143]. In this previous work, participants made decisions on whether or not to respond to the emails that were presented to them, resulting in thousands of email–response pairs. A total of 84 unique users participated in the study. For my study, I selected a subset of 250 emails from the full dataset due to resource constraints. This selection involved manually reviewing all emails and removing those that were too short, lacked meaningful content, or were too similar to others, while still ensuring a balanced mix of email types.

5.3.2 *Pilot Study*

The study was conducted in two stages: an initial pilot study followed by a full-scale user study. The pilot study was used to evaluate the study procedures and identify potential issues in the study design. During this phase, a small group of participants performed all the study activities: 1. evaluated two alternative summaries for each email and selected the one they felt best captured how they remembered the email 2. answered tone, sentiment, and intent rating questions. Observations from the pilot study provided insight into procedure design and question design. The pilot study informed the total number of emails included in the final study, the number of emails shown to each participant, and the optimal sequencing of the questions to minimize cognitive load.

5.3.3 *Full Study procedure*

Building on the insights gained from the pilot, I conducted a full study involving 150 participants who were recruited via Prolific. All participants were 18 years of age or older and resided in the United States. The survey was administered using Qualtrics, where participants reviewed an informed consent form before beginning the study. The study protocol received approval from the University of Washington Institutional Review Board (IRB). Each participant was compensated \$9 for completing the study, which was advertised as

a 45-minute session (equivalent to an hourly rate of approximately \$12). On average, the participants completed the study in about 40 minutes.

Each participant received 10 unique emails randomly selected from the dataset. Each email was reviewed by at least five different raters to ensure diversity of perspectives and improve the reliability and representativeness of the collected data. For each email, participants completed five tasks designed to capture the features necessary to compare the two categories of representation approaches described above. Below is an overview of the survey flow. The exact wording of all participant questions is included in Appendix A.

1. Evaluated email summaries produced by generative AI and select the summary that more closely aligns with how participants recalled the conversation.
2. Provided five keywords that best captured the core content or message of the email, during a phase in which the email was hidden to encourage recall-based responses.
3. Reported their perception of the tone of the email on a scale ranging from casual to formal.
4. Rated the overall sentiment of the email as negative, neutral, or positive.
5. Selected the primary intents of the email according to their perception of what was being communicated. These intent categories were adapted based on established frameworks in prior work on email intent modeling [72, 105, 115, 139]. See Table 5.1 for the questionnaire on email intents used in the study.

Table 5.1: Intent Categories and Descriptions Used in the Questionnaire

Intent Category	Description and Example
Request	Asking the recipient to take an action and/or provide information (e.g., password reset, answer a question)
Propose	Suggesting a joint action or agreement (e.g., proposing a meeting or collaboration)
Commit	Promising to take a future action (e.g., “I’ll send it tomorrow”)
Deliver	Providing information, recommendations, updates, or files (e.g., login alert, invoice, subscription confirmation)
Amend	Modifying a previous proposal, decision, or plan (e.g., changing meeting time or correcting info)
Refuse	Declining a request or offer (e.g., “I can’t attend,” “Sorry, not possible”)
Greet	Sending a social message (e.g., thank you, congratulations, greetings)
Remind	Sending a reminder about something (e.g., upcoming deadline, follow-up)
Resolve	Closing an issue, conflict, or question (e.g., confirming a problem is fixed, clarifying a misunderstanding, providing a final decision)

5.3.4 Data Format and Usage

The final dataset includes the following components, organized according to the two representation categories evaluated in this study which inform memory similarity and retrieval in the IBL models.

- **Human preferences for summaries:** Participants rated each summary on a 5-point scale in response to the question, “On a scale of 1 to 5, how well does this version of the email summary match how you recall the email?” (1 = not well at all,

5 = Extremely well). These ratings were used to identify high-quality summaries as lower-level text-based inputs for the IBL models.

- **Human-generated keyword sets:** Participants provided exactly five free-text keywords per email. These were intended to capture the most salient details participants remembered and were used to model recall-driven features in the IBL framework.
- **Annotated tone, sentiment, and intent labels:** Categorical labels derived from human responses:
 - *Tone* = {formal, casual, urgent}
 - *Sentiment* = {positive, neutral, negative}
 - *Intent* = {request, propose, commit, deliver, amend, inquire}

This dataset forms the foundation for evaluating which representation approach - low-level text-based features, such as summaries and keywords, or higher-level intent-based features, such as intent, tone, and sentiment - best supports human-aligned modeling of participant responses to phishing and benign emails.

5.3.5 *Summary-Based Representations*

Overview

To better align with how humans process and recall messages, I used model-generated summaries as lower-level text-based representations of each email. Summaries provide a compact alternative to full text, reducing the complexity and verbosity of lengthy emails while retaining the main informational content. This approach was motivated by the idea that people often recall the gist of an email rather than its full wording.

Data Collection

During the study, participants were shown two summaries for each email, generated by two transformer-based models and asked to rate each summary to determine how well it

matched how they would recall the email using a 1–5 scale. The original email text was visible while they evaluated the summaries to ensure that participants could make informed judgments. The summary with the highest average rating across participants was retained as the representative summary of that email.

The two summarization models evaluated in this study are:

- **T5 (Text-to-Text Transfer Transformer)** [93]: A transformer-based encoder-decoder model designed for a wide range of natural language tasks. T5 has demonstrated strong performance in generating high-quality summaries by learning to produce concise, meaningful text. In this study, the base variant of T5 was used to generate summaries that aimed to capture the key points of each email.
- **BART (Bidirectional and Auto-Regressive Transformers)** [69]: A denoising autoencoder that combines a BERT-like bidirectional encoder with a GPT-like autoregressive decoder. BART is pre-trained by applying various noising functions (e.g., token masking, sentence shuffling) and learning to reconstruct the original text, making it well-suited for sequence generation tasks like summarization.

In this study, I used automated summarization models to generate email summaries, as relying on multiple human-generated summaries could introduce inconsistency and noise. BART and T5 were evaluated to determine which model produced summaries most closely aligned with how people recalled the conversation. To ensure the outputs were reasonable and human-like, the automatically generated summaries were rated by human participants for quality and alignment with the original email content.

Instance Representation

Once the best summary was selected for each email, embeddings were generated using a sentence transformer model, specifically **all-MiniLM-L6-v2** [140]. This model, part of the MiniLM family of sentence transformers, is designed for efficient semantic representation with reduced computational overhead. It generates fixed-size 384-dimensional embeddings for any input sentence, ensuring that downstream models receive uniform-length vectors

regardless of the input text length. These embeddings capture key semantic relationships across sentences, providing a compact and meaningful representation of each email’s content [99].

By reducing irrelevant content, summarization was hypothesized to improve the focus of embeddings, thereby enabling the IBL model to retrieve memory traces that better reflect human recall patterns. This hypothesis is evaluated empirically in subsequent sections of this chapter.

By using embeddings of summaries instead of embeddings of the full email text, this representation approach aims to address two main challenges: (1) reducing the complexity and verbosity of lengthy email texts, and (2) potentially aligning better with human recall by providing shorter, more focused content than raw email text. In experiments described in Section 5.5, I compare IBL performance using summaries versus other representations to evaluate their impact on the model’s ability to approximate participant-like responses.

5.3.6 Recall-Based Keyword Representations

Overview

Another type of text-based representation explored in this dissertation is human-generated recall keywords. This approach is motivated by the observation that when recalling and searching for emails, people often rely on a few salient words or phrases rather than full sentences. Prior work also suggests that people compress information into keywords that capture the intent or tone of a message, rather than retaining its complete language [82, 64, 104]. This representation therefore provides a cognitively plausible and interpretable alternative to full-text or summary-based inputs, aligning with gist-based memory processes.

Data Collection

During the study, participants were asked to provide five keywords that represented how they would recall or search for each email in their inbox. Keywords were given in free-text format. For each email, keywords from all participants who viewed it were collated, resulting in a diverse pool of recall cues that captured different interpretations of the same

message.

To construct a final keyword representation, I used a two-phase selection strategy designed to balance human agreement and semantic diversity:

Phase 1: Frequency-Based Filtering Keywords mentioned by two or more raters were automatically included as reliable recall cues. If more than five such keywords existed, the five most frequent were chosen.

Phase 2: Diversity-Based Completion If fewer than five keywords remained after Phase 1, the set was completed by selecting additional keywords to maximize semantic diversity. This step was inspired by the Maximal Marginal Relevance (MMR) framework [25], which balances relevance with novelty.

The procedure was as follows:

- Sentence embeddings were generated for all candidate keywords using the `all-MiniLM-L6-v2` model from the SentenceTransformers library [140].
- Highly similar keywords (cosine similarity ≥ 0.85) were clustered, and only one representative was retained from each cluster to reduce redundancy.
- A **greedy selection algorithm** was then applied:
 - The process was seeded with the repeated keywords already included in Phase 1.
 - Iteratively, the next keyword chosen was the one with the lowest average cosine similarity to the already selected set, ensuring semantic diversity.
 - Because the initial seed can bias results, the process was repeated with each candidate as the starting seed. The final set chosen was the one with the lowest overall pairwise similarity across keywords.

This combined approach ensured that the final five keywords per email reflected both participant agreement and semantic diversity, aligning with the dual goals of cognitive plausibility and representational breadth. The five-term limit was also consistent with working

and benign emails.

5.3.7 *Tone, Sentiment, and Intent Label Aggregation*

Overview

In addition to collecting summaries and recall keywords, participants were asked to annotate each email with perceived tone, sentiment, and intent, capturing the primary communicative signals that humans rely on when interpreting conversations [61, 150]. These annotations provided higher-level abstract features for evaluating cognitively aligned representations.

Data Collection

Tone and Sentiment Participants selected one option for each dimension from a predefined set:

- **Tone:** Casual, Neutral, Formal
- **Sentiment:** Negative, Neutral, Positive

Each email was rated by multiple participants and the final tone and sentiment labels were assigned based on the majority vote. In cases of a tie between categories (for example, equal votes for 'positive', 'negative' and 'neutral'), the neutral option (for example, 'neutral' tone or 'neutral' sentiment) was assigned to avoid making strong assumptions not supported by participant consensus.

Intent Participants also selected one or more communicative intents for each email from a predefined taxonomy of email functions, including Request, Propose, Commit, Deliver, Amend, Refuse, Greet, Remind, and Resolve. These intent categories are grounded in prior research on email understanding and classification [72, 105, 115, 139]. The full set of intents used in this study is summarized in Table 5.1.

To simplify, the final intent representation was limited to a maximum of five intents per email. For each email, the intentions selected by all the participants who received it were collated and the frequency of each selected intent was calculated. The five most

frequently chosen intents per email were retained. If fewer than five intents were selected by raters, the final list included only the available high-agreement categories. This strategy enabled the model to consider the most salient intents while avoiding overly sparse or noisy representations that could undermine interpretability and alignment with human recall.

Figure 5.3 shows the distribution of chosen intents and sentiment labels across emails in the dataset.

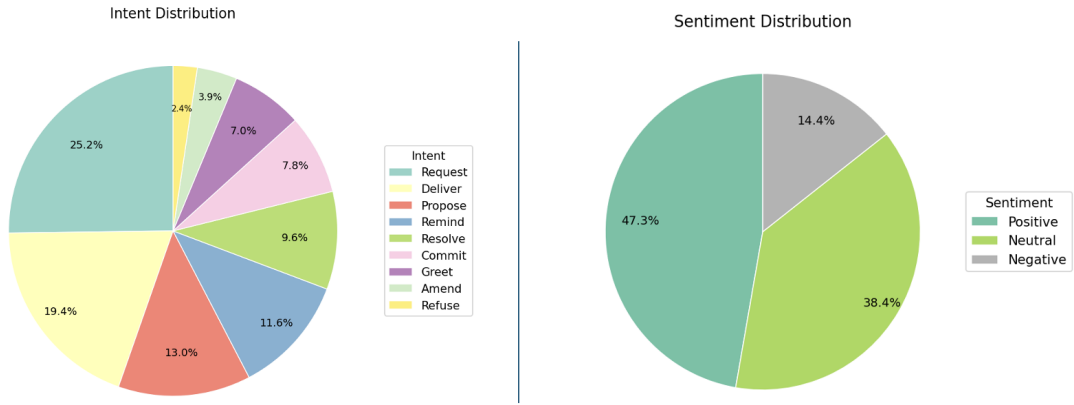


Figure 5.3: Distribution of chosen intents and sentiment labels across emails

Instance Representation

Tone, sentiment, and intent labels were aggregated into structured feature sets that served as higher-level abstract representations of each email. These categorical annotations provided the IBL model with cognitively meaningful cues, emphasizing the purpose and interpersonal qualities of the message rather than surface linguistic detail.

5.4 Method

Building on the instance representations introduced in Section 5.4.2, this section describes how they were implemented within the IBL framework to model human responses to phishing and benign emails.

For this study, I developed one IBL agent per participant to simulate responses to emails from the prior laboratory study [143]. Each agent was trained and tested on the

same emails as its human counterpart, using four variations of instance representations: (1) summaries, (2) keywords, (3) intent labels, and (4) a combined representation of intent, tone, and sentiment. These representations, ranging from low- to high-level features, were implemented as alternatives to the high-dimensional full-text embeddings used in prior work, aiming to better align with how people

5.4.1 IBL Framework and Activation Formula

As previously described, IBLT is a cognitive framework in which decisions are made by retrieving the most similar past instance from memory[48]. Each memory trace represents a prior experience, in this case, an email that a participant previously experienced and made decisions upon and new email experiences are compared to these traces using a activation function computed over shared features.

The retrieval strength of each memory instance is determined by the following activation equation:

$$A_{i,k,t} = \ln \left(\sum_{t_i=1}^t (t - t_i)^{-d} \right) + MP \cdot \sum_k \text{Sim}(v_k, c_k) + \sigma \cdot \ln \left(\frac{1 - \gamma_{i,k,t}}{\gamma_{i,k,t}} \right),$$

where:

- The first term represents the base-level activation, which decays as a function of both the time elapsed and the frequency of past encounters with the memory trace.
- The second term reflects partial matching, computed as the similarity (via cosine similarity) between the current cue vector v_k and the memory feature vector c_k .
- The third term introduces stochastic noise into the retrieval process, simulating the inherent variability in human memory access.

In this study, memory traces are defined using cognitively grounded features: participant-selected summaries or aggregated keyword vectors, or with tone, sentiment, and intent labels collected during the during. These cues were embedded to construct meaningful, human-aligned representations of each email.

5.4.2 Instance Representations

Building on the framework described above, I developed one IBL agent per participant to simulate their responses to phishing and benign emails from the prior laboratory study [143]. Each agent was trained and tested on the same emails as its human counterpart, modeling memory traces based only on that participant’s own experiences.

Four types of instance representations were implemented:

1. **Summaries:** LLM-generated summaries of each email.
2. **Keywords:** Self-reported keywords participants identified as useful for recalling each email.
3. **Intent:** Human-rated communicative intent labels.
4. **Intent + Tone + Sentiment:** Combined human-rated labels capturing the email’s perceived purpose, tone, and sentiment.

All representations, including categorical features, were embedded using the `all-MiniLM-L6-v2` sentence transformer to produce fixed-length vectors that preserved semantic relationships between features. Cosine similarity between embeddings was used for partial matching during retrieval. This setup enabled a direct comparison of representations ranging from compressed text (summaries, keywords) to higher-level abstract cues (intent, tone, sentiment), evaluating their impact on IBL model alignment with human responses.

5.5 Evaluation

5.5.1 Evaluation Objectives

The objective of this evaluation is to assess how effectively the different representations support the ability of IBL agents to predict human decision-making when responding to emails. Each IBL agent models the decision patterns of an individual participant and is evaluated based on its ability to reproduce that participant’s binary responses (e.g. respond/ignore) to previously unseen emails.

5.5.2 *Experimental Setup*

For each participant, an IBL agent was trained on that individual’s responses, constructing memory traces based on the emails they experienced during the previous study [143] using either lower-level text representations (summaries or keywords) or higher-level abstract cues (intent, tone, sentiment). The agent’s task was to predict the participant’s binary decision (“Response” or “Ignore”) on previously unseen emails.

Data for each rater was split into an 80/20 train/test partition, ensuring that test predictions were made only on emails not included in the agent’s training memory. This approach allows a direct assessment of how well different representations predict individual response on untrained email instances.

5.6 *Results*

5.6.1 *Overview and Objectives*

To compare the impact of different instance representation approaches on the performance of the IBL agent in predicting human response to emails, I report both traditional classification metrics - accuracy, precision, recall and F1 score - and metrics relevant to cognitive modeling such as average log likelihood [121], Akaike Information Criterion (AIC) [5] and Bayesian Information Criterion (BIC) [111] as indicators of model coplexity. While classification metrics quantify predictive performance, log-likelihood offers a more nuanced measure of how well each model captures an agent’s decision patterns across all emails. These complementary metrics enable direct comparisons in both predictive accuracy and alignment with individual raters’ behavioral profiles [121].

Additionally, I analyze log-likelihood heatmaps across key IBL parameters such as mismatch penalty and decay rates to show the trade-offs and optimal combinations of these parameters, and how model fit changes across different settings. By comparing agent-specific metrics, this analysis identifies which representations best capture individual variability in human decision-making patterns when responding to emails.

Overall, this evaluation seeks to determine which type of representation lower-level text features or higher-level abstract feature more effectively supports IBL modeling of human

decision-making in email response tasks.

I previously evaluated the IBL model using the full sentences of emails embedded as representations [79] (originally proposed by Xu and colleagues in [144]), which serves as the baseline for my analysis in the following sections of this chapter. This baseline achieved the following overall classification metrics across raters, providing a reference point for assessing the cognitive models developed in subsequent sections:

Accuracy	Precision	Recall	F1
60.0	55.4	48.3	49.9

5.6.2 Dataset and Evaluation Metrics

The evaluation in this chapter uses the dataset introduced earlier, consisting of 250 emails annotated by 150 participants. Each participant rated 10 unique emails across five dimensions: response decision, keywords, tone, sentiment, and intent. These annotations provide the foundation for training and evaluating IBL agents with both recall-oriented text representations (summaries, keywords) and higher-level abstract cues (intent, tone, sentiment).

Evaluation metrics include traditional classification measures: accuracy, precision, recall and F1 score to quantify predictive performance, as well as average log-likelihood to assess how well each model captures participant-specific decision patterns. Together, these metrics provide a comprehensive understanding of the effectiveness of IBL models with each representation.

In addition to variability in individual response tendencies, participants were exposed to differing distributions of email types. Figure 5.4 illustrates the number of phishing, primary, and promotional emails evaluated by each unique rater in the dataset I used in my study. While the majority of participants rated more primary and promotional emails, phishing emails were less prevalent overall. This imbalance increases the complexity of the modeling task, as each IBL agent must learn decision patterns from a unique set of examples with varying degrees of exposure to phishing content.

These differences in email type distribution highlight the necessity of agent-specific evaluation rather than assuming a homogeneous email stream across participants. Modeling

each rater individually ensures that the learning process and performance assessment accurately reflect the specific context and experience of each user, which is critical for capturing real-world variability in human decision-making.

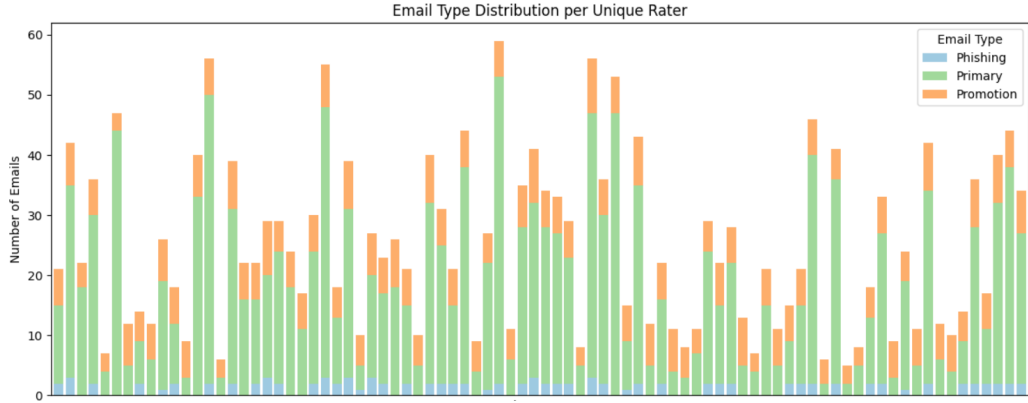


Figure 5.4: Distribution of email types (Phishing, Primary, Promotion) across individual raters.

5.6.3 Evaluation Metrics

To assess model performance, I report average classification metrics (accuracy, precision, recall, and F1-score) computed across all agents. These provide a baseline view of how well different representations support IBL agents to predict participants' responses. Beyond these metrics, I also used model fit measures relevant to cognitive modeling proposed by Stocco et al. [121]:

- **Log-Likelihood:** Quantifies how well an agent's predictions align with observed human responses, capturing the probability assigned to actual outcomes under the model. Following Stocco et al., it is computed as

$$\log L(m, \theta \mid \mathbf{x}) = \sum_{i=1}^N \log P(x_i \mid m, \theta),$$

where m denotes the participant-specific IBL agent, θ its parameters, $\mathbf{x}$ the sequence of human responses, and N the number of emails (trials) for that participant.

- **Akaike Information Criterion (AIC)**: Evaluates the trade-off between model fit and complexity, favoring models that achieve good predictive accuracy with fewer parameters:

$$AIC = 2k - 2 \ln L,$$

where k is the number of free parameters and $\ln L$ is the maximized log-likelihood.

- **Bayesian Information Criterion (BIC)**: Similar to AIC but applies a stricter penalty for model complexity, defined as

$$BIC = \ln(n) \cdot k - 2 \ln L,$$

where n is the number of data points (trials per agent).

Each metric is calculated individually for each IBL agent, representing the individualized model of each participant. The final results are reported as averages across agents to capture overall performance trends. This agent-level evaluation approach provides both population-level insights and the ability to analyze individual differences, aligning with best practices in personalized cognitive modeling.

5.7 Overall Performance Comparison

5.7.1 Comparative Evaluation of Cognitive Representations

To evaluate how the different representation approaches influence IBL agent performance, I compared four model variants spanning two categories of lower level and higher level representations.

Table 6.1 summarizes their performance across key metrics: accuracy, precision, recall, F1-score, log-likelihood, and complexity measures (AIC, BIC).

As shown in Table 6.1, models based on higher-level abstract cues generally outperformed lower-level text representations. Specifically, the **Intent-Sentiment-Tone-Based Model** achieved the highest performance overall, with accuracy, precision, and F1-score values indicating stronger alignment with participants' responses. This suggests that integrating

communicative intent with affective cues like tone and sentiment are closer to how people encode and recall email conversations from memory.

I also observed that the models showed differences in their probabilistic fit. The **Intent-Only Model** achieved the best (least negative) average log-likelihood (-5.18), indicating it produced better-calibrated probability estimates overall. In contrast, the **Keyword-Based Model** exhibited the worst log-likelihood (-7.98), despite outperforming the Summary-Based Model on classification metrics like precision and recall. Meanwhile, the **Intent-Sentiment-Tone Model** combined high accuracy (0.79) with a moderately negative log-likelihood (-6.12), suggesting a trade-off between strong prediction performance and probabilistic calibration.

Table 5.2: Comparison of Model Performance Across Instance Representations

Model	Accuracy	Precision	Recall	F1	AIC / BIC	Log-Likelihood
Summary-Based	0.65	0.40	0.67	0.54	10830.49 / 10840.93	-5.80
Keyword-Based	0.67	0.50	0.52	0.48	12614.69 / 12665.08	-7.98
Intent-Only	0.72	0.69	0.73	0.68	15243.88 / 15254.27	-5.18
Intent-Sentiment-Tone	0.79	0.62	0.75	0.66	15260.25 / 15275.35	-6.12

To better understand model fit at the individual level, I plotted the per-agent average log-likelihood (mean across each agent’s evaluation emails) for the four models (Figures 5.5–5.8). These plots reveal substantial variability in how well the models fit different participants, with some agents consistently achieving higher likelihoods than others. This variability—both across participants and across models—suggests that individual differences in decision-making may strongly influence model performance and motivated a deeper analysis to examine why certain participants were harder to model.

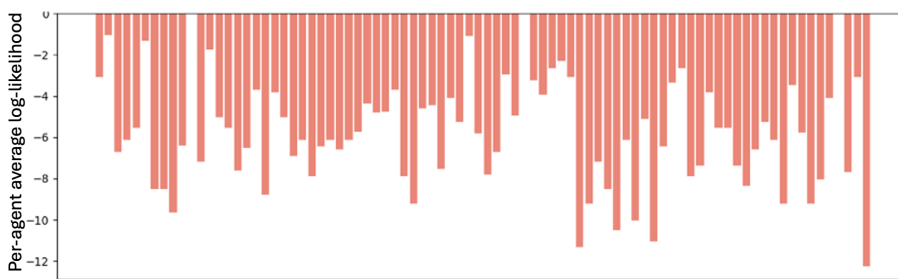


Figure 5.5: Average log-likelihood per agent for the Summary-Based Model.

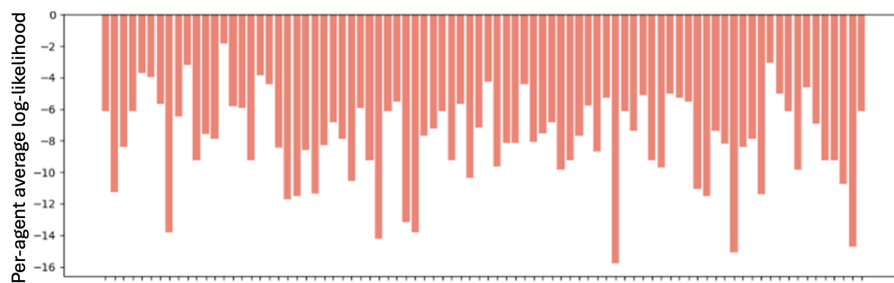


Figure 5.6: Average log-likelihood per agent for the Keyword-Based Model.

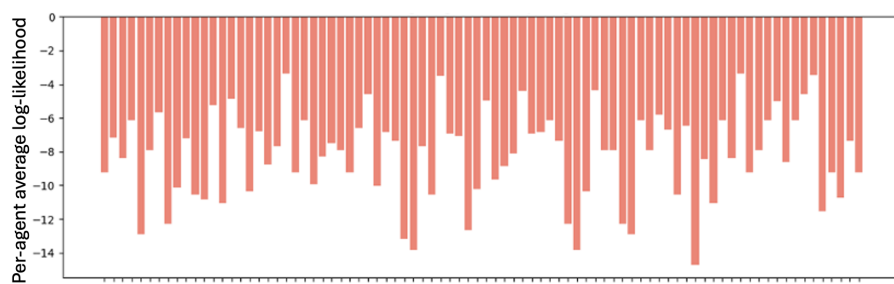


Figure 5.7: Average log-likelihood per agent for the Intent-Based Model.

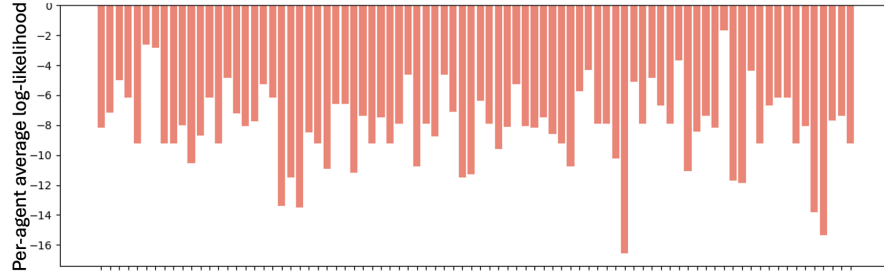


Figure 5.8: Average log-likelihood per agent for the Intent + Sentiment + Tone Model.

5.7.2 Clustering Participants by Modeling Performance

The substantial variability in agent log-likelihoods across participants motivated a deeper analysis to understand whether certain individuals were consistently harder to model and why. To explore this, I first quantified **participant consistency**, defined as the proportion of a participant’s responses that matched the majority decision for the same emails. Higher consistency values indicate participants whose judgments align more closely with the collective consensus, whereas lower values reflect more random and inconsistent decision-making.

Next, I clustered participants based on their overall modeling performance. Each cognitive agent was represented by a 20-dimensional feature vector consisting of accuracy, precision, recall, F1-score, and average log-likelihood across the four evaluated models. All metrics were standardized using z-scores to ensure equal contribution across dimensions. KMeans clustering with $k = 3$ was applied to partition participants, producing three distinct groups. A principal component analysis (PCA) projection of these clusters is shown in Figure 5.9.

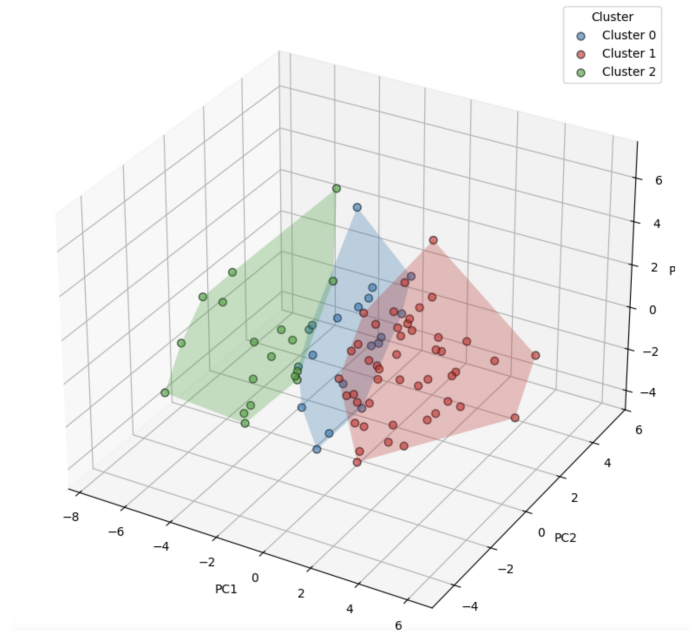


Figure 5.9: PCA projection of participant performance profiles. Clusters represent distinct groups of raters based on model fit metrics.

The three clusters correspond to meaningful differences in participant behavior:

- **Cluster 0:** High-consistency raters (average agreement $\approx 70\%$)
- **Cluster 2:** Moderate-consistency raters (average agreement $\approx 60\%$)
- **Cluster 1:** Low-consistency raters (average agreement $\approx 53\%$)

To examine how model performance varied across these groups, I compared average F1 scores for each model by cluster (Figure 5.10). Models relying on compressed text representations (**Summary-Based** and **Keyword-Based**) showed step drops in F1 scores for low-consistency participants (Cluster 1), suggesting that these simpler representations are less robust for raters with less predictable decision patterns. In contrast, the **Intent** and **Intent + Sentiment + Tone** models maintained more stable F1 scores across clusters, indicating greater resilience to participant inconsistency.

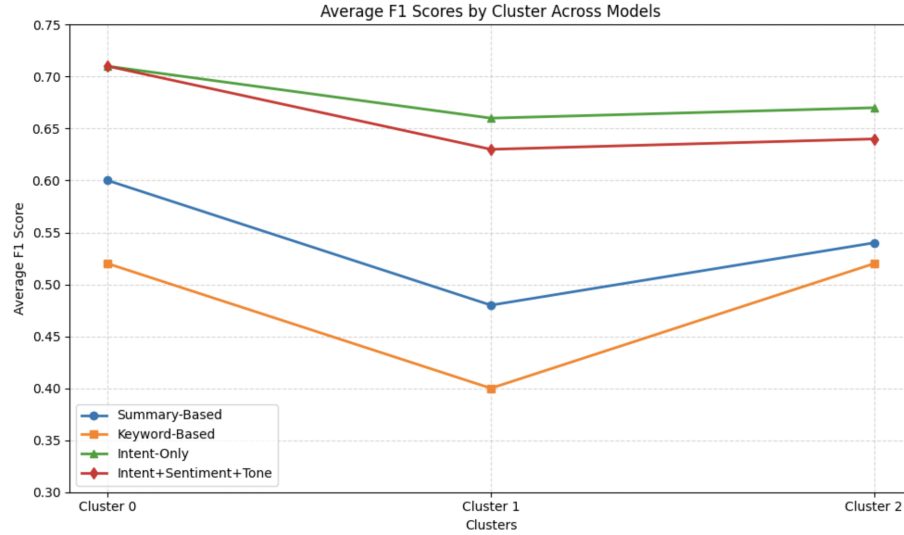


Figure 5.10: Average F1 scores by cluster for each model.

Similar patterns emerged in the log-likelihood analysis: agents for high-consistency raters achieved less negative likelihoods (indicating better fit), while agents for low-consistency raters had substantially lower likelihoods, particularly for the simpler models. This convergence between accuracy-based metrics and likelihood evidence suggests that participant consistency is a strong driver of model fit.

Overall, these results indicate that raters with stable, majority-aligned decision patterns are easier to model, especially with simpler text-based representations. In contrast, participants with more inconsistent or idiosyncratic behavior expose the limitations of these approaches, while cue-rich models (intent, sentiment, tone) demonstrate greater robustness across diverse rater profiles.

5.8 Individual Agent Analysis

To better understand how rater consistency affects model fitting, I examined log-likelihood heatmaps of individual agents representative of Cluster 1 (least consistent) and Cluster 0 (most consistent). Figure 5.11 shows heatmaps for an agent from Cluster 1 (least consistent), while Figure 5.12 shows those for an agent from Cluster 0 (most consistent).

For the inconsistent agent in Cluster 1, summary- and keyword-based models exhibit diffuse or fragmented dark regions in their heatmaps, lacking well-localized optima. This instability suggests that small parameter adjustments cause large fluctuations in log-likelihood, underscoring the difficulty of modeling raters whose decisions deviate unpredictably. In contrast, the intent-based models, including intent-only and intent-sentiment-tone, still show relatively structured optima even for this inconsistent agent—indicating that these representations are more robust to inconsistency.

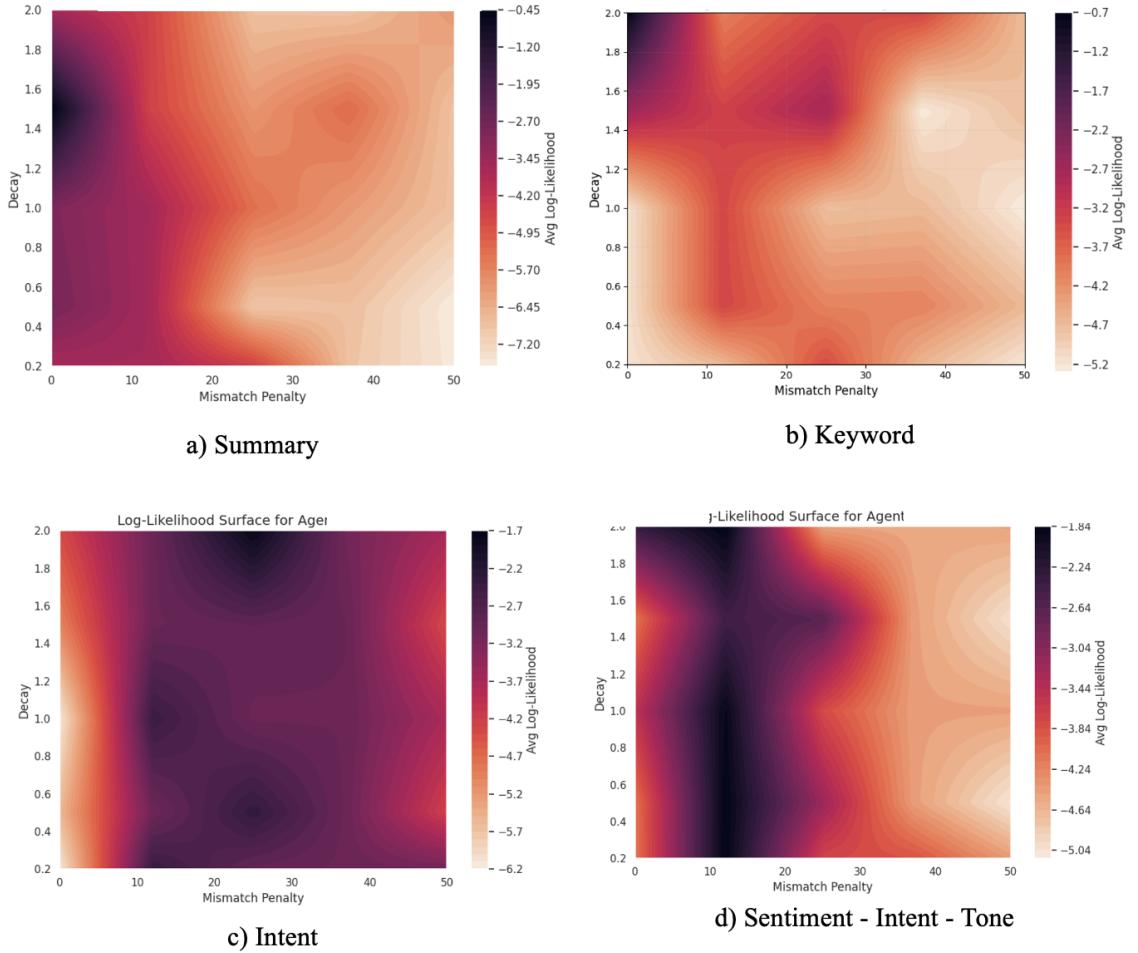


Figure 5.11: Log-likelihood heatmaps for an agent from Cluster 1 (inconsistent rater).

For the consistent agent in Cluster 0, heatmaps from summary- and keyword-based mod-

els display distinct, well-localized optima across mismatch penalty and decay parameters. These stable optima suggest that for consistent raters, even text based models can reliably identify parameter configurations that capture their response patterns.

These qualitative assessments reinforce earlier quantitative findings: intent-based models maintain more stable fits despite rater inconsistency, while summary- and keyword-based models degrade sharply when faced with unpredictable response patterns.

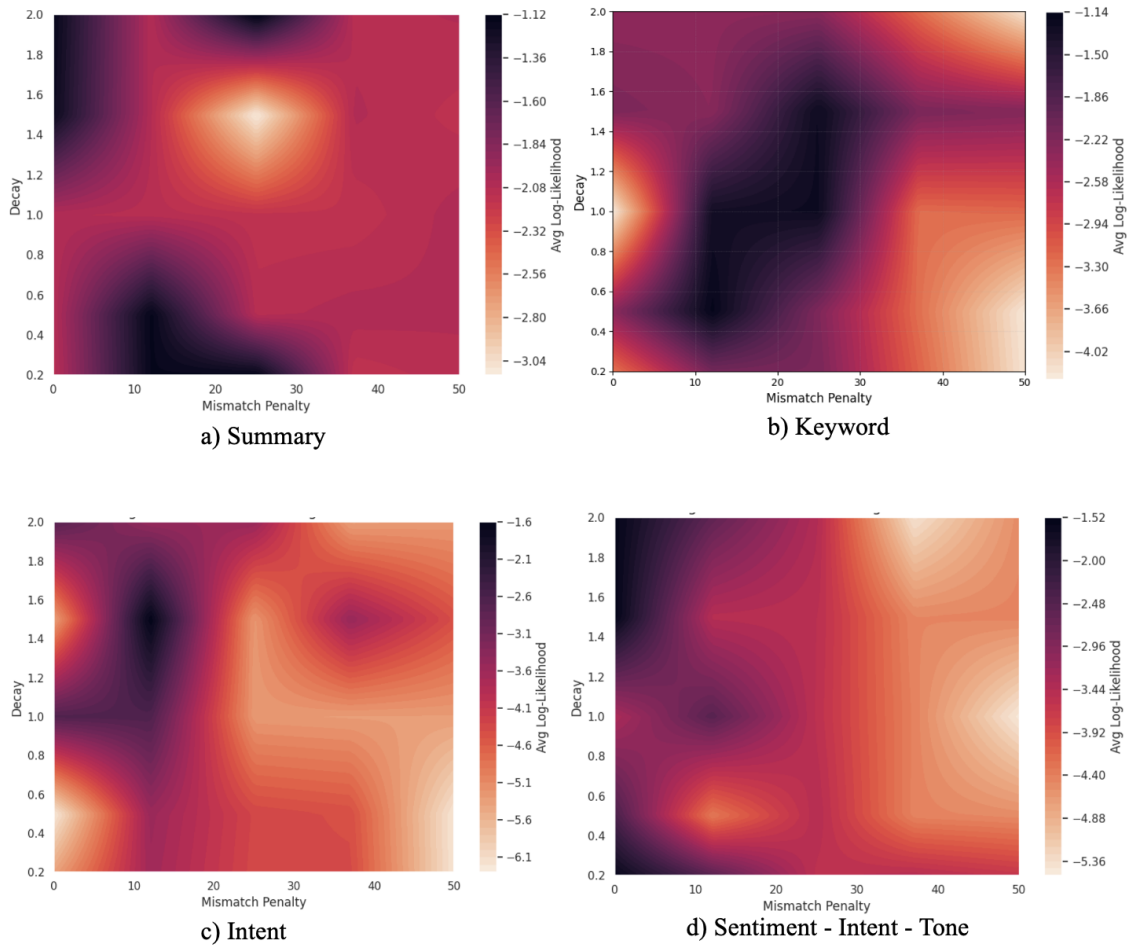


Figure 5.12: Log-likelihood heatmaps for an agent from Cluster 0 (consistent rater).

5.9 Prediction Agreement by Email Type

To investigate how model predictions align with human responses across Phishing, Primary, and Promotion emails, I visualized correctness patterns for each model. In each plot, green indicates correct predictions and red indicates errors, organized by email type.

For the Summary-Based model, overall prediction accuracy was 68% for Phishing, 60% for Primary, and 65% for Promotion emails (Figure 5.13).



Figure 5.13: Prediction agreement for the Summary-Based model across email types.

For the Keyword-Based model, overall prediction accuracy was 70% for Phishing, 65% for Primary, and 66% for Promotion emails (Figure 5.14).

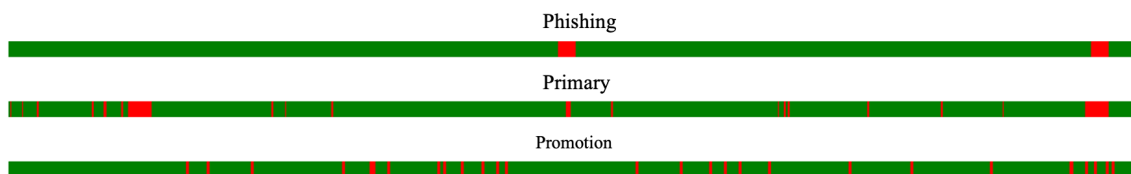


Figure 5.14: Prediction agreement for the Keyword-Based across email types.

Errors in the Summary-Based model are broadly distributed, particularly among Primary and Promotional emails, underscoring the difficulty of distinguishing nuanced categories using summaries alone. In contrast, the Keyword-Based model shows slightly improved performance for phishing emails, suggesting that participant-generated keywords provide more targeted cues for identifying deceptive content.

For the Intent-Only model, overall prediction accuracy was 75% for Phishing, 70% for Primary, and 71% for Promotion emails (Figure 5.15).

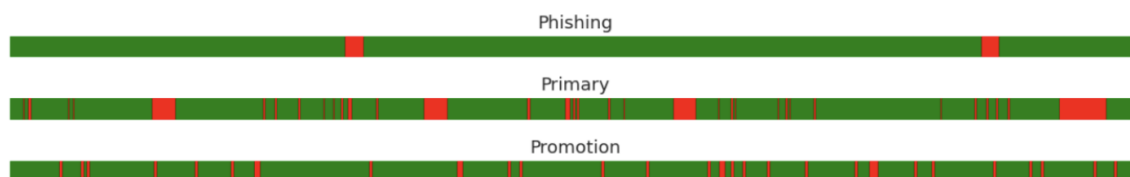


Figure 5.15: Prediction agreement for the Intent-Only model across email types.

For the Intent-Sentiment-Tone model, overall prediction accuracy was 82% for Phishing, 78% for Primary, and 77% for Promotion emails (Figure 5.16).

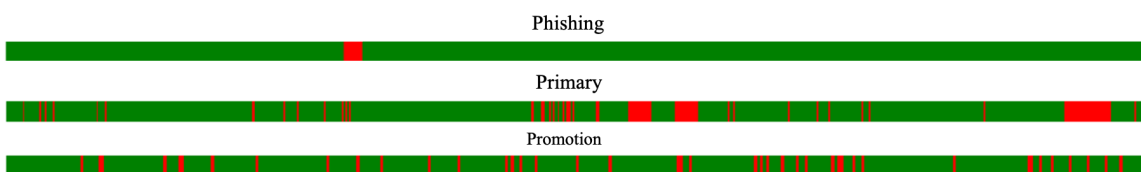


Figure 5.16: Prediction agreement for the Intent-Sentiment-Tone model across email types.

In general these two intent based models show a better performance over phishing category with significantly higher accuracy.

Deeper analysis of misclassified emails revealed that phishing messages most frequently missed by the intent model but correctly detected by intent-sentiment-intent models, were dominated by intents like *Request* (approximately 20% of misclassified phishing), followed by *Deliver* (17%) and *Propose* (13%). This pattern highlights that these intentions are highly context sensitive and require auxiliary cues such as sentiment or tone for an accurate interpretation. Integrating cues beyond intent significantly improves the model’s ability to capture subtle patterns of phishing tactics. For example, a “Request” intent could reflect a benign scenario (e.g., “Please confirm your attendance”) or a malicious one (e.g., “Urgent: Confirm your bank details”), but intent alone cannot distinguish these possibilities. The

addition of sentiment and tone helps disambiguate these cases, enabling the model to better align with human interpretations of suspicious requests.

5.10 Discussion and Future Work

This chapter investigated how different types of instance representations influence the performance and interpretability of IBL models in simulating human responses to emails. Specifically, I systematically compared low-level text embeddings with higher-level, cognitively enriched features, including inferred intent, tone, and sentiment. The results showed that models using intent-based and other high-level embeddings performed significantly better than those relying on raw text. This finding suggests that incorporating cognitively meaningful features can enhance the alignment between cognitive models and human decision-making processes in the context of email evaluation.

One interpretation of this finding is that people may encode and recall email communications not at the level of raw text but rather based on inferred intent and the associated actions [82, 64, 104]. Although email text is typically unstructured and varies widely in length, tone, and writing style, the underlying function of emails often follows common communicative patterns, such as requesting information, delegating tasks, confirming events, or providing updates [31, 106, 72]. Intents therefore introduce a latent structure that is not always captured by text-level embeddings. Intent-based representations may better reflect this underlying structure by focusing on what the email is trying to achieve, rather than on the specific words or style used to convey the message [139, 115]. People may adopt a similar strategy when recalling email communications—inferring the gist or purpose of a message and storing this abstracted representation—potentially explaining why intent-based features improved model performance.

Another explanation relates to how cognitive models, such as IBL, function. IBL relies on similarity-based retrieval from memory, and such similarity-based functions tend to perform better when input instances are represented using higher-order, semantically meaningful features like intents. Because intent-based representations are more structured and less variable than raw text embeddings, they create a more coherent memory space. This coherence facilitates more accurate retrieval of relevant prior experiences, resulting in

decisions that better mirror human memory and judgment processes. This interpretation is further supported by my cluster analysis, which showed that intent-based representations performed well across both consistent and inconsistent rater clusters, suggesting that the improved performance may stem from a more coherent underlying memory structure.

Together, these findings have important implications for future cognitive modeling of phishing decision-making. They suggest that incorporating techniques such as intent modeling and sentiment modeling can be valuable for automatically generating more human-aligned instance representations. By abstracting away from the variability of raw text and focusing on meaningful cognitive cues, such models can achieve greater fidelity in simulating how humans process, recall, and act on email communications.

Despite these promising findings, several limitations highlight opportunities for future research. First, the heatmap analyses revealed substantial heterogeneity across raters, indicating that while some individuals benefit significantly from enriched cues, others show minimal improvement. However, interpreting log-likelihood surfaces at the individual level remains qualitative and subjective. Future research could address this by incorporating clustering techniques or computational surface analysis metrics to systematically quantify topography. Such methods would enable automated identification of rater types based on their log-likelihood landscapes, providing a more rigorous and scalable approach to understanding individual differences in decision-making.

Second, the summary, keyword, intent, and contextual cue features were not collected or generated uniformly. While summaries were generated by models, intent, tone, and sentiment labels were annotated by raters and may vary in reliability or granularity. These inconsistencies may have introduced biases affecting predictive accuracy. Future work should aim to improve representation consistency, either through more rigorous annotation protocols, improved inter-rater agreement measures, or by leveraging fine-tuned models explicitly aligned with domain-specific cues to reduce noise and enhance feature quality.

Finally, there is potential bias in feature selection. Although the inclusion of keywords, tone, and sentiment was grounded in cognitive theories of memory and attention, these choices may have unintentionally biased the models toward emails with emotional or overtly spam-like characteristics, potentially limiting generalizability to more subtle or professional

phishing attempts. Future research should expand the feature space to include structural cues (e.g., email formatting, sender identity), behavioral indicators (e.g., click-through probabilities), or even physiological measures (e.g., eye-tracking) to better approximate human perception and decision-making in realistic email scenarios.

Overall, this chapter advances our understanding of how cognitively enriched representations impact the performance of cognitive models in simulating human responses to emails. By demonstrating the value of intent-based and contextual features, it lays the groundwork for developing more human-aligned, adaptive modeling approaches. The next phase of this research builds on these insights, exploring how these enriched representations can be systematically integrated into cognitive modeling frameworks to improve generalizability and interpretability.

Chapter 6

EXPLORING ACT-R BASED PHISHING DECISION MAKING MODELS

6.1 *Motivation*

Up to this point, cognitive modeling in this dissertation has primarily focused on similarity-based retrieval mechanisms (IBL), particularly through Instance-Based Learning. Although IBL provides a powerful framework for modeling memory-driven decision making, it represents only one aspect of the broader cognitive processes that underpin how people could be processing and responding to email conversations.

According to dual-process theories of decision making [60, 136], human cognition operates through two complementary systems: System 1, which is fast, intuitive and automatic, and System 2, which is slower, more deliberate, and analytical. In the context of email evaluation, System 1 processes may dominate routine and familiar tasks, enabling individuals to make quick judgments based on learned heuristics or rules, whereas System 2 processes may be engaged when interpreting novel, complex, or ambiguous content.

This distinction highlights a critical gap in existing IBL-based models: they capture similarity-based retrieval processes but do not fully explain the rapid, rule-driven judgments that often occur early in email evaluation. IBL has been proposed as a framework for studying intuitive decision-making within cognitive architectures [123], yet it does not account for rule-based intuitive processes, which are considered a core form of intuitive reasoning [32]. These rule-based shortcuts may precede instance retrieval in familiar or low-effort decision contexts, and addressing them is a central focus of this chapter.

To address this gap, this chapter explores rule-based cognitive modeling within the Adaptive Control of Thought–Rational (ACT-R) architecture. ACT-R provides a well-established framework for representing knowledge in the form of production rules [8], allowing models to simulate the fast, heuristic-driven decisions characteristic of System 1 processing. By

incorporating production rules alongside retrieval-based mechanisms, this approach aims to capture a more comprehensive range of cognitive strategies used by individuals when responding to phishing and benign emails.

6.2 Introduction and Background

Adaptive Control of Thought–Rational (ACT-R) is a cognitive architecture developed by Anderson and colleagues, designed to simulate and explain human cognitive processes [122]. ACT-R provides a unified framework for modeling how knowledge is represented, processed, and used in the human mind. The architecture combines symbolic components such as production rules for explicit reasoning with subsymbolic mechanisms that govern memory activation and retrieval, enabling it to capture both deliberate and automatic aspects of cognition. ACT-R has been applied to model a wide range of tasks, from simple reaction time experiments to complex domains like language comprehension, decision-making, and human-computer interaction [9].

ACT-R consists of multiple specialized modules that correspond to distinct aspects of human cognition, including the declarative memory module (responsible for factual knowledge), the procedural memory module (for learned rules and skills), the goal module (which manages current tasks) and perceptual-motor modules (handling sensory input and motor actions) [20, 9]. These modules communicate through buffers, which act as interfaces between the memory systems and the central production system. Each buffer can hold a single chunk at a time, where chunks represent structured units of information such as concepts, objects, or goals [20, 147].

The procedural system in ACT-R operates by evaluating production rules –IF—THEN statements – that match patterns across the buffers[9]. When the conditions of a production rule match the current buffer contents, the rule can fire and execute actions such as updating goals, issuing retrievals, or modifying buffer states. This cycle of matching, selecting, and firing productions forms the core of ACT-R’s decision-making process.

Declarative memory in ACT-R stores chunks that are retrieved based on their activation level. The activation A_i of a chunk i is determined by a combination of base-level activation B_i , spreading activation SA_i from currently active chunks, and noise ε sampled from a

logistic distribution:

$$A_i = B_i + SA_i + \varepsilon \quad (6.1)$$

The base-level activation reflects the frequency and recency of use:

$$B_i = \ln \sum_{k=1}^n (t_k)^{-d} \quad (6.2)$$

where t_k is the time since the k -th retrieval and d is the decay parameter.

Spreading activation SA_i is computed based on associative links from the goal and other active chunks:

$$SA_i = \sum_{j \in \text{cues}} W_j \cdot S_{ji} \quad (6.3)$$

where W_j is the attentional weight of the source chunk and S_{ji} is the associative strength from chunk j to chunk i .

Production rules are selected through conflict resolution. When multiple productions are applicable, ACT-R selects the one with the highest utility U_i , computed as:

$$U_i = P_i \cdot G - C_i \quad (6.4)$$

Here, P_i is the estimated probability that production will achieve the goal, G is the value of the goal, and C_i is the expected price of taking that action—time, effort, or risk—so the higher the price, the larger the expected benefit ($P_i \cdot G$) must be for that production to be chosen. This utility-based mechanism allows ACT-R to adaptively learn and refine strategies over time [44].

In this chapter, I introduce ACT-R and its production rule system as a modeling framework for phishing decision making. Although my overarching aim of modeling human decision-making remains consistent with earlier chapters, this section explores an alternative approach that involves the use of production rules alongside similarity-based decision making. Specifically, I test whether symbolic production rules derived using decision trees could further enhance the performance of cognitive models of phishing decision-making. This bridges data-driven modeling with cognitive theory, grounding production rules in empirically observed behavioral patterns [92].

Production rules in ACT-R are typically constructed using one of three approaches:

- The first is **expert-defined rules**, where the rules are manually crafted based on theoretical insights or insights from experiments on heuristics that people could be using while making decisions in a context [9]. This approach offers high interpretability but limited scalability to complex, real-world scenarios.
- The second approach is **model-internal learning**, in which ACT-R dynamically adjusts rule utilities or compiles new rules through experience-based mechanisms [9]. While this method captures learning over time, it often requires extensive interaction traces and detailed task simulations.
- The third approach is **data-driven rule induction**, where behavioral data are analyzed to extract condition-action patterns, often through interpretable machine learning models like decision trees[77]. These patterns can then be translated into production rules, bridging empirical observations with symbolic cognitive modeling [101, 74].

The method presented in this chapter belongs to the third category of data-driven approaches. By translating paths derived from decision tree models into ACT-R production rules, my goal is to develop interpretable, modular, and cognitively plausible models of phishing decision making. These production rules capture empirical patterns observed in individual-level behavioral data, specifically decisions to respond or ignore emails. This approach enables the grounding of symbolic cognitive models in real-world data, bridging the gap between statistical machine learning and cognitive theory. Ultimately, this chapter contributes a hybrid method that integrates data-driven insights with the symbolic reasoning capabilities of ACT-R, demonstrating a practical pathway for building cognitive models that are informed by empirical behavior and rooted in established cognitive architectures.

6.3 Objectives

The objective of this chapter is to develop an interpretable rule-based cognitive model using the ACT-R architecture to simulate human decision-making when responding to phishing emails. Unlike the similarity-based approaches explored in earlier chapters, this model encodes decision strategies as symbolic IF–THEN production rules informed by empirical

data on email intent and related cues. By integrating these rules with ACT-R’s declarative and procedural memory systems, the chapter aims to create individualized agents capable of capturing both shared and rater-specific decision patterns in a cognitively plausible way.

6.4 Method

6.4.1 Dataset and Feature Representation

This study uses the same dataset introduced in the previous chapter, consisting of 250 emails annotated by human participants. Each rater provided a binary response (*Respond* or *Ignore*) for every email and annotated several perceptual attributes, including:

- **Tone:** *Casual*, *Formal*, or *Neutral*.
- **Sentiment:** *Positive*, *Neutral*, or *Negative*.
- **Intent:** Categories such as *Request*, *Deliver*, or *Remind*, with multiple selections allowed per email.

In addition, participants rated the perceived **importance** of each email on a 0–10 scale. For modeling, I averaged these scores across the raters and discretized them into two bins: `importance_low` for scores ≤ 5 and `importance_high` for scores > 5 . Two additional binary features were included:

- `link`: indicating whether the email contained any hyperlinks.
- `spam`: indicating whether participants labeled the email as spam.

Together, these perceptual judgments and content-based cues provide a rich set of symbolic features for modeling email evaluation. Each feature was transformed into binary indicators compatible with ACT-R’s chunk-based representation. For example, tone ratings were encoded as `tone_casual`, `tone_neutral`, or `tone_formal`; sentiment ratings as `sentiment_positive`, `sentiment_neutral`, or `sentiment_negative`; and intent selections as binary flags for each observed category (e.g., `intent_request`, `intent_deliver`).

Finally, the **human response label** (*Respond* or *Ignore*) served as a target variable for modeling. These structured symbolic representations allowed the derivation of production rules for ACT-R and supported the subsequent evaluation of the models' ability to represent participant decision patterns.

6.4.2 Decision Tree Rule Induction

To derive interpretable symbolic rules suitable for integration into ACT-R, I trained a decision tree classifier using the engineered binary features derived from human annotations and email content. The objective was to learn mappings from structured email representations to human response decisions (**Respond** vs. **Ignore**). Decision trees were chosen because they produce human-readable IF-THEN rules that can be directly translated into symbolic ACT-R production rules, bridging empirical data with cognitively plausible decision processes.

Model Setup: The modeling pipeline was implemented in Python using `scikit-learn`. Starting with the data set of annotated emails, I expanded the multi-label categorical features (`intent`, `tone`, and `sentiment`) into binary feature vectors through a custom multi-hot encoding process. Each possible label (e.g., `intent.deliver`, `tone.formal`, `sentiment.positive`) was represented as a separate binary feature indicating whether that label was present in the email. The numerical importance ratings were scored with a threshold of 5, producing a binary flag (`importance.high`) to reflect the high perceived relevance. Binary content features, including `link` and `spam`, were also included to capture the presence of hyperlinks and the likelihood of the email being perceived as spam.

The input matrix X consisted of these binary features, while the target variable y was the human response label, which was binarized using `LabelEncoder` to enable model training. A `DecisionTreeClassifier` was trained with a maximum depth of 5 (`max.depth=5`). Each decision path from root to leaf can be interpreted as a symbolic rule encoding a sequence of feature-based conditions leading to a predicted behavioral outcome. These paths serve as candidate ACT-R production rules grounded in empirical behavioral data and structured representations of email content and perceptual cues.

Rule Extraction: Each path from the root of the decision tree to a leaf node was interpreted as a conjunctive IF condition, representing the sequence of feature-based decisions along that path. The predicted class label at the leaf node (**Respond** or **Ignore**) was treated as the corresponding THEN action. For example, one learned rule extracted from the trained decision tree was:

Example Rule

IF importance_high = 1 AND tone_formal = 1 AND sentiment_positive = 0 THEN
Respond

Each such rule provides a human-interpretable description of decision strategies captured by the model, enabling direct translation into ACT-R production rules. Because each decision path involves a limited number of binary features (typically 3–5), the resulting rules remain both interpretable and cognitively plausible.

In total, I selected a small set of representative rules covering a diverse range of perceptual and content-based email conditions, with particular emphasis on those that occurred frequently or reflected psychologically meaningful distinctions (e.g., importance, tone, intent).

Rule Selection Criteria: While the decision tree produced a large number of possible decision paths, only a carefully selected subset was retained for ACT-R modeling. Rule selection followed best practices in cognitive modeling, grounded in both empirical precedent and theoretical constraints. Specifically, rules were selected according to the following criteria:

1. **Frequency of Rule Activation:** Rules that applied to a substantial portion of the dataset were prioritized, aligning with ACT-R’s base-level activation theory, which emphasizes the role of experience in strengthening knowledge traces [74].
2. **Cognitive Plausibility:** Rules were restricted to a maximum of 3–5 conjunctive conditions, consistent with ACT-R’s assumptions about bounded working memory capacity and limits on the complexity of production rules [9].

This selection process reflects a hybrid methodology that combines data-driven learning (via decision tree induction) with cognitively grounded model construction. The resulting symbolic rules serve as a bridge between observed human decisions and ACT-R's procedural reasoning mechanisms.

Below are Eight exemplar rules extracted from the trained tree, each represented in natural language and boxed for clarity:

Rule 1: Important + Formal $\rightarrow$ Respond

IF importance_high = 1 **AND** tone_formal = 1 **AND** sentiment_positive = 0
AND intent_amend = 0 **THEN** Respond

Interpretation: Emails rated as important and formal, but not overly positive in tone, were typically responded to. These may be interpreted as professional but serious communications that demand attention.

Rule 2: Friendly Greeting $\rightarrow$ Respond

IF importance_high = 0 **AND** intent_greet = 1 **THEN** Respond

Interpretation: Even emails not marked as important were likely to receive a response if they contained a greeting, suggesting a socially driven response pattern.

Rule 3: Casual + Positive $\rightarrow$ Respond

IF importance_high = 0 **AND** tone_casual = 1 **AND** sentiment_positive = 1
THEN Respond

Interpretation: Light, casual messages with a positive tone were often responded to even when not marked as important, possibly reflecting low-cost social engagement.

Rule 4: Formal + Amend + Positive → Ignore

IF importance_high = 1 **AND** tone_formal = 1 **AND** sentiment_positive = 1
AND intent_amend = 1 **THEN** Ignore

Interpretation: Polished, amendment-related messages—even when important—were more likely to be ignored, perhaps perceived as non-urgent or purely informational.

Rule 5: Important + Negative Tone → Respond

IF importance_high = 1 **AND** sentiment_negative = 1 **AND** tone_formal = 0
THEN Respond

Interpretation: Negative messages marked as important but informal may have conveyed urgency or concern, prompting responses.

Rule 6: Casual Greetings → Ignore

IF importance_high = 0 **AND** intent_greet = 1 **THEN** Ignore

Interpretation: Casual greetings in unimportant emails were often ignored, likely perceived as friendly but non-essential.

Rule 7: Positive Spammy Proposals → Ignore

IF importance_high = 0 **AND** sentiment_positive = 1 **AND** spam = 1 **AND**
intent_propose = 1 **AND** link = 1 **THEN** Ignore

Interpretation: Overly positive emails with proposals and links flagged as spam were likely seen as promotional or suspicious, leading to non-response.

Rule 8: Spammy Amendments in Formal Tone → Ignore

IF importance_high = 0 **AND** intent_amend = 1 **AND** spam = 1 **AND**
tone_formal = 1 **THEN** Ignore

Interpretation: Messages combining formal amendments and spam signals were often dismissed as boilerplate or manipulative content.

6.4.3 Production Rule Translation into ACT-R

Once symbolic IF–THEN rules were extracted from the decision tree, the next step was to translate these rules into ACT-R production rules. In ACT-R, production rules represent procedural knowledge in the form of conditional behavior: each production inspects the contents of buffers (e.g., goal, retrieval, visual) and, if its conditions match, performs actions such as modifying the goal, initiating a retrieval, or issuing motor/perceptual commands [9, 20].

Mapping Structure: In the implemented model, extracted IF–THEN rules were directly translated into ACT-R productions that match features in the goal buffer. When the model’s goal chunk contained feature values exactly matching the rule conditions, the corresponding production would fire immediately, updating the goal’s state and specifying the predicted decision (**respond** or **ignore**).

A representative decision rule such as:

Sample Rule

```
IF importance_high = 1 AND tone_formal = 1 AND sentiment_positive = 0
AND intent_amend = 0 THEN Respond
```

was implemented in ACT-R syntax as:

```
(p respond-to-formal-negative
=goal>      state decide
            importance_high 1
            tone_formal 1
            sentiment_positive 0
            intent_amend 0
==>
```

```
=goal>      state done
           decision respond)
```

In this modeling approach, if no production matched the current goal chunk, the model initiated a retrieval from declarative memory to locate a previously experienced chunk with features similar to the current email’s cues. A separate fallback production then used the retrieved chunk’s decision to update the goal, completing the decision process.

This two-stage mechanism—direct rule matching followed by retrieval-based decision-making—was designed to simulate both explicit, rule-driven strategies and memory-based reasoning, reflecting ACT-R’s hybrid symbolic and subsymbolic capabilities while staying faithful to its architectural principles.

Chunk and Buffer Configuration: In the implemented model, each email was encoded as a chunk stored in ACT-R’s declarative memory. These chunks contained binary attributes corresponding to the engineered feature set (e.g., `importance_high`, `tone_formal`, `sentiment_positive`, etc.), maintaining consistency with the symbolic representations used during decision tree training.

The goal buffer held a chunk representing the current decision context, initialized with the features of the email under consideration and a `state` slot set to `decide`. The retrieval buffer was configured to search declarative memory for previously experienced email chunks whose features closely matched those in the goal chunk when no production rule fired immediately. This setup enabled the model to simulate decision-making either through direct symbolic rule matching or by retrieving relevant past experiences—capturing both explicit and memory-based processes within ACT-R’s architecture.

Buffer definitions and chunk types were configured following the conventions of the ACT-R cognitive architecture, implemented in Python using PyACT-R. For example:

```
model.chunktype("email", [
    "importance_high", "action", "meeting", "link",
    "tone_formal", "tone_neutral", "tone_casual",
    "sentiment_positive", "sentiment_neutral", "sentiment_negative",
```

```

    "intent_request", "intent_deliver", "intent_remind"
])
model.set_goal("goal")
model.set_retrieval("retrieval")

```

These configurations defined the structure of email chunks and established the `goal` and `retrieval` buffers used for modeling decision processes. Although inspired by the official ACT-R architecture, this implementation was conducted entirely in Python.

In summary, I translated decision tree-derived symbolic rules into ACT-R productions that directly map email features to response behaviors. Each email was encoded as a chunk with binary attributes, and the ACT-R model used these chunks to match the conditions specified in the production rules. Although the ACT-R conflict resolution system was engaged in every cycle, this implementation usually resulted in only one rule matching per decision state. Utility learning was not applied. This approach enabled a clear and interpretable modeling of how key perceptual and content features influenced human decision patterns in email responses.

Retrieval Dynamics and Cognitive Plausibility: In this implementation, ACT-R’s declarative memory retrieval operates through activation-based nearest-neighbor search, allowing the model to retrieve chunks that best match current cues even when exact feature matches are absent. Chunk activation reflects both the recency and frequency of prior retrievals, consistent with ACT-R’s base-level learning mechanism. This ensures that chunks used more frequently or more recently become more accessible, capturing key dynamics of human memory. Although the current model uses the default parameters of ACT-R for partial matching and spreading activation, this set-up effectively enables the system to handle different patterns of email features by retrieving the closest matching chunks in most cases. This initial configuration establishes a foundation for future work, such as tuning mismatch penalties or incorporating associative links to model richer memory interactions.

6.4.4 *Simulation Setup and Evaluation Procedure*

To assess the performance of the ACT-R models, a simulation-based evaluation framework was implemented for each individual rater. Each ACT-R agent was initialized with the same set of symbolic production rules but trained on rater-specific experience data - chunks derived from that rater’s own email responses. During the evaluation, each agent was presented with structured feature representations of emails from the test set, and its predicted responses were recorded. These predictions were then compared with the actual responses provided by the corresponding rater to compute accuracy, precision, recall, and other performance metrics. This approach allowed the same cognitive model architecture to capture individual differences leveraging each rater’s unique interaction history.

Model Initialization: For each unique rater in the data set, a separate instance of the ACT-R model was created. Although all models shared the same chunk type definitions and production rules derived from the decision tree, each maintained an individualized declarative memory containing only the chunks representing emails previously encountered by that specific rater. This individualized memory allowed the model to simulate how the unique experiences of each rater influenced their response decisions during evaluation.

Memory Encoding: Each email associated with a given rater was encoded as a chunk containing binary attributes corresponding to the engineered features (e.g., `importance_high`, `meeting`, `intent_request`, etc.). These chunks were added to the model’s declarative memory using ACT-R’s chunk construction functions in the Python implementation.

Performance Metrics: After simulating all trials for each rater, standard evaluation metrics, including Accuracy, Precision, Recall, and F1 were computed.

Example Outputs

To illustrate how the ACT-R models generate decisions during simulation, I provide examples of two types of evaluation cases: matched and unmatched trials.

Matched Cases: These are trials in which the test email’s features directly satisfied the conditions of one of the extracted symbolic production rules, leading to an immediate rule match and firing of the corresponding ACT-R production. Figure 6.1 shows a sample output trace of this process, including conflict resolution, rule selection, rule firing, and goal state modification.

```
(0, 'PROCEDURAL', 'CONFLICT RESOLUTION')
(0, 'PROCEDURAL', 'RULE SELECTED: rule8')
(0.05, 'PROCEDURAL', 'RULE FIRED: rule8')
(0.05, 'goal', 'MODIFIED')
(0.05, 'PROCEDURAL', 'CONFLICT RESOLUTION')
(0.05, 'PROCEDURAL', 'NO RULE FOUND')

===== Result =====
Predicted decision: ignore
```

Figure 6.1: Example of a matched rule firing directly.

Unmatched Cases: These occur when none of the symbolic production rules exactly match the cues of the test email. In such trials, the ACT-R model defaults to retrieving the closest matching chunk from declarative memory to generate a decision. Figure 6.2 shows a trace of this retrieval-based process, where no rule was initially found, triggering a retrieval that ultimately guided the predicted response.

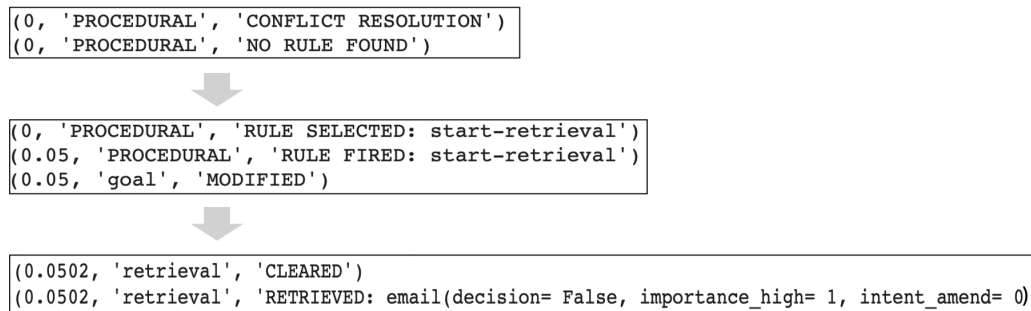


Figure 6.2: Example of retrieval-based decision due to no matching rule.

These examples illustrate the two-stage decision-making process implemented in the

ACT-R model: direct production firing when rules match, and fallback retrieval when they do not. This combination allows the model to simulate both symbolic decision strategies and memory-based reasoning.

Overall, the last row of Table 6.1 shows the ACT-R model’s average performance metrics across all raters compared to the previous approaches:

Table 6.1: Comparison of Model Performance Across Instance Representations

Model	Accuracy	Precision	Recall	F1
Summary-Based	0.65	0.40	0.67	0.54
Keyword-Based	0.67	0.50	0.52	0.48
Intent-Only	0.72	0.69	0.73	0.68
Intent-Sentiment-Tone	0.79	0.62	0.75	0.66
ACT-R and Production Rules	0.73	0.68	0.88	0.73

These results demonstrate that, even with a limited symbolic rule set, the ACT-R models can simulate individual differences in decision behavior with high recall and competitive precision.

6.5 Discussion

This study demonstrated the feasibility and potential of using ACT-R production rules to model human decision-making when responding to emails. By combining ACT-R’s rule-based reasoning with similarity-based retrieval, this approach provided an interpretable alternative to the purely similarity-driven IBL framework examined in earlier chapters.

One key advantage of the rule-based approach is its interpretability: decision-making logic is encoded in explicit IF–THEN rules, making it possible to trace and understand the reasoning behind model predictions. This transparency is particularly valuable for applications where explaining or auditing model behavior is critical. Furthermore, this approach aligns with dual-process theory [60, 136] by incorporating both fast, rule-driven

decisions and slower, retrieval-based responses, offering a more nuanced account of human decision-making processes.

Performance analysis revealed an important behavioral tendency: all models exhibited a bias toward predicting “respond,” which increased recall. This bias likely stems from the class imbalance in the dataset, where “ignore” (typically phishing) emails were substantially less frequent than “respond” emails. While this bias improved sensitivity, it underscores the need for future testing on more balanced datasets to better evaluate the model’s robustness.

6.5.1 Limitations

Despite these contributions, several limitations highlight opportunities for future research. First, the production rules used here were derived from decision trees trained on a single dataset collected within a controlled study. While effective for modeling observed behavior in this context, these rules may not generalize well to other datasets, user populations, or real-world email environments. Future work should explore training on multiple datasets or integrating domain knowledge to develop rules that capture more universal cognitive strategies.

Second, the current model uses fixed rule priorities. Although ACT-R’s similarity-based retrieval supports fallback decisions when no rule matches, the system does not yet assign or update rule utilities based on experience. Incorporating dynamic utility learning would allow the model to adapt over time, reflecting how humans refine decision strategies through feedback and practice.

In summary, this chapter represents an initial step toward symbolic, interpretable cognitive modeling of email decision-making. By integrating production rules with ACT-R’s memory mechanisms, this work moves beyond purely retrieval-based approaches, providing a framework that is both cognitively grounded and practically interpretable. Future research can build on this foundation by improving generalizability, implementing adaptive rule learning, and testing on more diverse datasets to fully realize the potential of rule-based cognitive models for understanding decision-making in complex communication environments.

Chapter 7

CONCLUSION

In the face of increasing cybersecurity threats, particularly phishing attacks, there is an urgent need to train users to recognize and respond appropriately to deceptive emails. Importantly, we need autonomous methods that provide frequent and personalized testing and training to improve phishing awareness. However, effective training and simulation require models that can closely mimic how humans actually process, recall past phishing trainings, and make more accurate decisions on real phishing threats. Despite advances in machine learning and natural language processing, there remains a gap: Existing anti-phishing systems fail to model the cognitive mechanisms that shape how people remember and recall phishing training while making decisions on real threats. This dissertation addresses that gap by developing cognitively grounded models that align more closely with human memory and decision-making processes.

The research began by integrating previously developed Instance-Based Learning (IBL) models into a simulated phishing study environment. These models were intended to automate human-like responses to phishing emails by drawing on prior experiences stored as memory traces. However, early evaluations revealed a key limitation: models using raw or full-text embeddings lacked the fidelity to accurately model and predict how humans recall and respond to emails. I hypothesized that the models in previous work were not performing well because they were designed with input instances that capture the selective and abstract nature of human memory, which prioritizes encoding and recalling high-level cues over exact text.

Motivated by this insight, I investigated approaches that leveraged higher-order, cognitively meaningful features — such as perceived intent, emotional tone, and key ideas — as instance representations within the IBL models. Through a pipeline that learned these representations from human-labeled data, I demonstrated that it is indeed possible to bridge

the gap between raw email text and the kinds of email features people actually encode and recall. This marked the first step toward building models that more closely align with the content and structure of human memory.

Motivated by the limitations of using raw email text as input, I next explored a privileged learning approach that used human-labeled survey data as lower-dimensional, higher-order features within the IBL framework. These survey-derived labels, which capture how participants classified email topics, were used to train neural networks capable of predicting similar labels for new unlabeled emails, providing a pathway to generalize these features beyond the original dataset. Although this approach improved model performance and demonstrated the feasibility of bridging raw text with structured and human-informed features, it also exposed critical challenges. In particular, the quality and stability of these representations varied depending on participant consistency, email type, and the degree to which the learned features generalized across datasets.

These findings motivated a broader and more systematic investigation of which types of representations are most effective in modeling human email decisions within the IBL framework. In the next phase of my research, I evaluated two main categories of representations:

1. **Lower-level text-based features:** including full-text embeddings, summary embeddings, and keyword-based embeddings.
2. **Higher-level abstract features:** including perceived intent, sentiment, and emotional tone.

Across participants and email types, I found that models using higher-level features, especially intent, consistently outperformed those relying on lower-level text-based representations. These higher-order features allowed the models to generalize across diverse emails that shared communicative goals, even when their surface language differed.

Furthermore, I introduced a participant-clustering analysis that revealed how representation effectiveness varies with individual consistency: abstract features provided greater robustness to variability, while lower-level features were more sensitive to inconsistent rater behavior. This analysis underscored the importance of accounting for individual differences

when evaluating model performance and highlighted the benefits of higher-level representations for improving model stability across diverse participants.

In addition to similarity-based modeling, I explored symbolic modeling as a complementary approach. I translated decision tree-derived rules into interpretable production rules within the ACT-R cognitive architecture. This allowed me to simulate human decisions using explicit, rule-based mechanisms grounded in cognitive theory. These symbolic models demonstrated competitive predictive performance while offering the advantage of transparency and interpretability. Importantly, they revealed that many human-like decisions could be captured through a small set of generalized rules, especially when those rules operated over abstract features like intent or urgency.

Together, the IBL and ACT-R models provide two complementary lenses for understanding and simulating human responses to phishing emails. The IBL models capture the nuance of memory-based retrieval and feature similarity, while the ACT-R models illustrate structured decision strategies based on cognitive rules.

7.1 Contributions

This dissertation makes several key contributions:

- Identifies and addresses the gap in existing knowledge on how to represent text as input instances to train cognitive models of human decision making.
- Introduces a privileged learning pipeline for learning higher-level email features aligned with human interpretation.
- Establishes a unified evaluation framework to compare diverse representations within IBL models, revealing that abstract features outperform surface-level text features.
- Demonstrates how symbolic ACT-R models can simulate human decisions using interpretable production rules derived from data.
- Provides empirical evidence that cognitively grounded features improve both model performance and interpretability.

7.2 *Future Directions*

Some important paths remain for future exploration. Future work should explore scaling to more naturalistic data by testing these models on real-world email corpora—including organizational and multilingual datasets—to assess robustness in realistic settings and under domain shift, while maintaining appropriate privacy protections. Future work should also examine multimodal behavioral cues by incorporating additional human signals—such as eye-tracking, response latency, and mouse movement—to provide richer context for modeling attention and hesitation and to clarify when messages feel convincing or risky. Another promising direction is hybrid cognitive architectures: combining symbolic (ACT-R) and subsymbolic (IBL) components in an integrated framework to leverage structured rules alongside similarity-based memory, potentially yielding more flexible and adaptive decision-making. Finally, future work should develop adaptive and personalized modeling to build user-specific cognitive models that update over time as people encounter new emails or learn new phishing tactics, with attention to calibration, transparency, and changes in attacker behavior.

7.3 *Closing Summary*

In summary, this dissertation advances our understanding of how to design cognitive models specifically for phishing decision making. It however also can be generalized to other decision making situations involving text processing. It demonstrates that by aligning model representations with how humans interpret and recall information, we can build more accurate, interpretable, and human-like models. These findings pave the way for developing cognitively informed models and designing anti-phishing training systems that more accurately reflect human decision-making processes in email interactions and other complex communication contexts.

BIBLIOGRAPHY

- [1] Yasaman D Abbasi, Noam Ben-Asher, Cleotilde Gonzalez, Debarun Kar, Don Morrison, Nicole Sintov, and Milind Tambe. Know your adversary: insights for a better adversarial behavioral model. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 38, 2016.
- [2] R. Abdulsalam and S. Alzahrani. Phishing: An overview of the threat landscape. *International Journal of Information Security*, 19(5):453–467, 2020.
- [3] Alessandro Ecclesie Agazzi. Phishing and spear phishing: examples in cyber espionage and techniques to protect against them, 2020.
- [4] Palvi Aggarwal, Omkar Thakoor, Shahin Jabbari, Edward A Cranford, Christian Lebiere, Milind Tambe, and Cleotilde Gonzalez. Designing effective masking strategies for cyberdefense through human experimentation and cognitive models. *Computers & Security*, 117:102671, 2022.
- [5] Hirotugu Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723, 2003.
- [6] A. Aleroud and L. Zhou. Phishing environments, techniques, and countermeasures: A survey. *Computers & Security*, 68:160–196, 2017.
- [7] Mohd Muddassir Ali and Nur Faezah Mohd Zaharon. Phishing—a cyber fraud: The types, implications and governance. *International Journal of Educational Reform*, 33(1):101–121, 2024.
- [8] John R Anderson. Production systems and the act-r theory. In *Rules of the mind*, pages 1–15. Psychology Press, 2014.
- [9] John R Anderson, Daniel Bothell, Michael D Byrne, Scott Douglass, Christian Lebiere, and Yulin Qin. An integrated theory of the mind. *Psychological review*, 111(4):1036, 2004.
- [10] R. Azuma, M. Daily, and C. Furmanski. A review of time critical decision making models and human cognitive processes. In *2006 IEEE aerospace conference*, pages 9–pp, 2006.
- [11] Alan Baddeley. Working memory: An overview. *Working memory and education*, pages 1–31, 2006.

- [12] V. Baisa. Chunk-based language model and machine translation. In *Proceedings of the 20th International Conference on Computational Linguistics*, 2005.
- [13] M. Bakhshaei, R. Kargari, and M. Ranjbar. The rise of spear phishing: A comprehensive review. *Computer Science Review*, 45:100–125, 2023.
- [14] Colin Bannard and Chris Callison-Burch. Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL’05)*, pages 597–604, 2005.
- [15] Abdul Basit, Maham Zafar, Xuan Liu, Abdul Rehman Javed, Zunera Jalil, and Kashif Kifayat. A comprehensive survey of ai-enabled phishing attacks detection techniques. *Telecommunication Systems*, 76:139–154, 2021.
- [16] Armin Bergholz, Katja De Beer, Stefan Glahn, Marie-Francine Moens, Gerhard Paaß, and Siegfried Strobel. A new perspective on phishing detection. In *Proceedings of the 2010 International Conference on Computational Linguistics*, page 1086–1094, 2010.
- [17] Santosh Kumar Birthriya, Priyanka Ahlawat, and Ankit Kumar Jain. Detection and prevention of spear phishing attacks: A comprehensive survey. *Computers & Security*, page 104317, 2025.
- [18] Jim Blythe, Aaron Botello, Joseph Sutton, David Mazzocco, Jerry Lin, Marc Spraragen, and Mike Zyda. Testing cyber security with simulated humans. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 25, pages 1622–1627, 2011.
- [19] Michael Bossetta. The weaponization of social media: Spear phishing and cyber attacks on democracy. *Journal of International Affairs*, 71:97–106, 2018.
- [20] Dan Bothell. Act-r 7.0+ reference manual. <http://act-r.psy.cmu.edu/actr7.x/reference-manual.pdf>, 2007. Carnegie Mellon University.
- [21] Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*, 2015.
- [22] J Bransford and J Franks. The abstraction of linguistic ideas cognitive psychology, 1971.
- [23] John D. Bransford and Marcia K. Johnson. Contextual prerequisites for understanding: Some investigations of comprehension and recall. *Journal of Verbal Learning and Verbal Behavior*, 11(6):717–726, 1972.

- [24] Caitlin I Canfield, Baruch Fischhoff, and Ashley Davis. Suspicion and trust of online information: The role of communication in shaping judgment. *Journal of Experimental Psychology: Applied*, 22(2):159–173, 2016.
- [25] Jaime Carbonell and Jade Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336, 1998.
- [26] David Carmel, Yi Chang, Hongbo Deng, and Jian-Yun Nie. Future directions of query understanding. In *Query Understanding for Search Engines*, pages 205–224. Springer, 2020.
- [27] Rich Caruana and Y Li. Detecting intent types in email conversations. In *Proceedings of the 2011 Conference on Email and Anti-Spam*, 2011.
- [28] Ritu Chadha, Thomas Bowen, Cho-Yu J Chiang, Yitzchak M Gottlieb, Alex Poylisher, Angello Sapello, Constantin Serban, Shridatt Sugrim, Gary Walther, Lisa M Marvel, et al. Cybervan: A cyber security virtual assured network testbed. In *MILCOM 2016-2016 IEEE Military Communications Conference*, pages 1125–1130. IEEE, 2016.
- [29] Tommy Chin, Kaiqi Xiong, and Chengbin Hu. Phishlimiter: A phishing detection and mitigation approach using software-defined networking. *IEEE Access*, 6:42516–42531, 2018.
- [30] François Chollet et al. Keras. <https://keras.io>, 2015.
- [31] William W Cohen, Vitor R Carvalho, and Tom M Mitchell. Learning to classify email into “speech acts”. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 309–316, 2004.
- [32] Brendan Conway-Smith and Robert L West. System-1 and system-2 realized within the common model of cognition. In *AAAI 2022 fall symposium*, 2022.
- [33] F. I. M. Craik and R. S. Lockhart. Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, 11(6):671–684, 1972.
- [34] Edward A Cranford, Christian Lebiere, Prashanth Rajivan, Palvi Aggarwal, and Cleotilde Gonzalez. Modeling cognitive dynamics in end-user response to phishing emails. In *Proceedings of the 17th ICCM*, 2019.
- [35] D Dang-Pham and S Pittayachawan. Toward understanding the impact of emotion on phishing susceptibility: A review of the literature. *Information & Computer Security*, 2017.

- [36] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.
- [37] Rachna Dhamija, J Doug Tygar, and Marti Hearst. Why phishing works. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 581–590, 2006.
- [38] Julie S Downs, Mandy B Holbrook, and Lorrie Faith Cranor. Decision strategies and susceptibility to phishing. In *Proceedings of the second symposium on Usable privacy and security*, pages 79–90, 2006.
- [39] Mark Dredze, Hanna Wallach, David Puller, and Fernando Pereira. Entity-based email classification. In *Proceedings of the Conference on Email and Anti-Spam*, 2008.
- [40] S. Egelman, Y. F. Cheng, and L. F. Cranor. Usability, security, and privacy: The role of user education in phishing attacks. *Computers Security*, 44:126–138, 2014.
- [41] Keelan Evans, Alsharif Abuadbba, Tingmin Wu, Kristen Moore, Mohiuddin Ahmed, Ganna Pogrebna, Surya Nepal, and Mike Johnstone. Raider: Reinforcement-aided spear phishing detector. In *Network and System Security: 16th International Conference, NSS 2022, Denarau Island, Fiji, December 9–12, 2022, Proceedings*, pages 23–50. Springer, 2022.
- [42] Siamak Farshidi, Kiyan Rezaee, Sara Mazaheri, Amir Hossein Rahimi, Ali Dadashzadeh, Morteza Ziabakhsh, Sadegh Eskandari, and Slinger Jansen. Understanding user intent modeling for conversational recommender systems: a systematic literature review. *User Modeling and User-Adapted Interaction*, pages 1–64, 2024.
- [43] Federal Bureau of Investigation. Internet crime report 2024. https://www.ic3.gov/AnnualReport/Reports/2024_IC3Report.pdf, 2024. [Online].
- [44] Wai-Tat Fu and John R. Anderson. From recurrent choice to skill learning: A reinforcement-learning model. *Journal of Experimental Psychology: General*, 135(2):184–206, 2006.
- [45] J. Glasgow. Integrating cognitive science and phishing research. *Journal of Cybersecurity Studies*, 45(2):101–115, 2023.
- [46] Sanjay Goel, Kevin Williams, and Ersin Dincelli. Got phished? internet security and human vulnerability. *Journal of the Association for Information Systems*, 18(1):2, 2017.

- [47] Cleotilde Gonzalez, Noam Ben-Asher, Alessandro Oltramari, and Christian Lebiere. Cognition and technology. In *Cyber defense and situational awareness*, pages 93–117. Springer, 2014.
- [48] Cleotilde Gonzalez, Javier F Lerch, and Christian Lebiere. Instance-based learning in dynamic decision making. *Cognitive Science*, 27(4):591–635, 2003.
- [49] M. D. González, F. Martínez, and J. Romero. Spear phishing and its impact on organizations. *Journal of Cybersecurity Research*, 5(2):85–95, 2022.
- [50] Wayne D Gray, Michael J Schoelles, and Christopher W Myers. Task constraints in applied cognitive modeling. In *Integrated Models of Cognitive Systems*, pages 62–75. Oxford University Press, 2006.
- [51] A.-P.W. Group and et al. Phishing activity trends report 2023 q4. Technical report, 2024. Retrieved from https://docs.apwg.org/reports/apwg_trends_report_q4_2023.pdf.
- [52] Rakesh Gupta, Nitin Koul, and et al. Intent-aware contextual recommendation using hierarchical rnn with attention. In *Proceedings of the 12th ACM Conference on Recommender Systems*, pages 423–427, 2018.
- [53] Z. M. Hakim, N. C. Ebner, D. S. Oliveira, S. J. Getz, B. E. Levin, T. Lin, and R. C. Wilson. The phishing email suspicion test (pest): A lab-based task for evaluating the cognitive mechanisms of phishing detection. *Behavior Research Methods*, 53:1342–1352, 2021.
- [54] Z.M. Hakim, N.C. Ebner, D.S. Oliveira, S.J. Getz, B.E. Levin, T. Lin, K. Lloyd, V.T. Lai, M.D. Grilli, and R.C. Wilson. The phishing email suspicion test (pest): A lab-based task for evaluating the cognitive mechanisms of phishing detection. *Behavior Research Methods*, pages 1–11, 2020.
- [55] Jeffrey T Hancock, Lauren E Curry, Saurabh Goorha, and Michael Woodworth. Communicative goals and deception: An analysis of interpersonal deception theory. *Discourse Processes*, 45(6):401–426, 2007.
- [56] Fred Heiding, Simon Lermen, Andrew Kao, Bruce Schneier, and Arun Vishwanath. Evaluating large language models’ capability to launch fully automated spear phishing campaigns: Validated on human subjects. *arXiv preprint arXiv:2412.00586*, 2024.
- [57] Grant Ho, Asaf Cidon, Lior Gavish, Marco Schweighauser, Vern Paxson, Stefan Savage, Geoffrey M Voelker, and David Wagner. Detecting and characterizing lateral phishing at scale. In *28th USENIX security symposium (USENIX security 19)*, pages 1273–1290, 2019.

- [58] Jason Hong. The state of phishing attacks. *Communications of the ACM*, 55(1):74–81, 2012.
- [59] Infosec Institute. A brief history of spear phishing. <https://www.infosecinstitute.com/resources/phishing/a-brief-history-of-spear-phishing/>, 2019. Accessed: 2025-07-21.
- [60] Daniel Kahneman. *Thinking, fast and slow*. macmillan, 2011.
- [61] Amir Khosrowpour, Yandong Ma, Ahmed Hassan Awadallah, and Ryen W. White. Pragmatic intent classification for email response suggestion. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7295–7305, 2020.
- [62] Yejin Kim and Alice Oh. Automatic keyword extraction using frequency and position information. In *Proceedings of the 3rd International Conference on Knowledge Capture*, pages 53–60. ACM, 2009.
- [63] W. Kintsch. The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, 95:163–182, 1988.
- [64] W. Kintsch and T. A. van Dijk. Toward a model of text comprehension and production. *Psychological Review*, 85:363–394, 1978.
- [65] Walter Kintsch. *Comprehension: A paradigm for cognition*. Cambridge University Press, 1998.
- [66] Y. Kumaraguru, A. Rhee, L. F. Acquisti, J. Cranor, J. Hong, and E. Nunge. 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*, pages 905–914, 2007.
- [67] C Lebiere. A blending process for aggregate retrievals. In *Proceedings of the 6th ACT-R Workshop*, 1999.
- [68] Wendy G Lehnert. Plot units and narrative summarization. In *Cognitive Science*, volume 5, pages 293–331, 1981.
- [69] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, 2020.

- [70] Bohan Li, Hao Zhou, Junxian He, Mingxuan Yang, Yiming Wang, and Lei Li. When text embedding meets large language model. *arXiv*, 2024. Notably, embedding methods “still have limitations in capturing complex semantic and syntactic structures”:contentReference[oaicite:6]index=6.
- [71] Xiangyu Li, Zhu Su, Jian Shen, Qing Li, Shuzi Niu, Yan Zhao, and Minlie Huang. A survey on dialogue systems: Recent advances and new frontiers. *ACM Transactions on Information Systems*, 38(4):1–40, 2020.
- [72] Chung-Ching Lin, Dong Kang, Michael Gamon, and Patrick Pantel. Actionable email intent modeling with reparametrized rnns. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.
- [73] G. W. Lindsay. Attention in psychology, neuroscience, and machine learning. *Frontiers in Computational Neuroscience*, 14:29, 2020.
- [74] Marsha C Lovett, Larry Z Daily, and Lynne M Reder. A source activation theory of working memory: Cross-task prediction of performance in act-r. *Cognitive Systems Research*, 1(2):99–118, 2000.
- [75] Zhan Luo and Min Zhu. Recurrent neural networks with mixed hierarchical structures for natural language processing. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2021.
- [76] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge: MIT Press, 2008.
- [77] Julian N Marewski and Katja Mehlhorn. Using the act-r architecture to specify 39 quantitative process models of decision making. *Judgment and Decision making*, 6(6):439–519, 2011.
- [78] James L. McClelland, Bruce L. McNaughton, and Randall C. O’Reilly. Why there are complementary learning systems in the hippocampus and neocortex. *Psychological Review*, 102(3):419–457, 1995.
- [79] Elaheh Mehrabi and Prashanth Rajivan. Neural network-driven cognitive models of phishing decisions: Evaluating a privileged learning framework for instance representation. Paper presented at Virtual MathPsych/ICCM 2025, June. Via <https://mathpsych.org/presentation/1984>, June 2025.
- [80] Joseph Menn. Exclusive: Russia used facebook to try to spy on macron campaign – sources. *Reuters*, July 2017. Accessed: 2025-07-02.
- [81] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

- [82] George A Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, 63(2):81, 1956.
- [83] A. Momeni, K. Tatwawadi, and U. Stanford. Title of the paper is not provided. *Journal/Conference name is not provided*, page Pages are not provided, 2022.
- [84] D Morrison and Cleotilde Gonzalez. Pyibl, a python implementation of iblt.
- [85] A. Newell and H. Simon. *Human Problem Solving*. Prentice-Hall, Oxford, England, 1972.
- [86] Antti Oulasvirta and Jan Blom. Motivations in personalisation behaviour. *Interacting with computers*, 20(1):1–16, 2008.
- [87] Bimal Parmar. Protecting against spear-phishing. *Computer Fraud & Security*, 2012(1):8–11, 2012.
- [88] M. Patterson. Cognitive mechanisms of instance-based learning: A review. *Cognitive Psychology Review*, 12(4):200–218, 2023.
- [89] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.
- [90] R. F. Port. How are words stored in memory? beyond phones and phonemes. *Language and Cognitive Processes*, 22(4):477–509, 2007.
- [91] Proofpoint. The 2022 cost of phishing study, 2022.
- [92] J. Ross Quinlan. Simplifying decision trees. *International Journal of Man-Machine Studies*, 27(3):221–234, 1987.
- [93] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.
- [94] P Rajivan, V Shankaranarayanan, and NJ Cooke. Cybercog: a synthetic task environment for studies of cyber situation awareness. In *Presentation and Proceedings of 10th International Conference on Naturalistic Decision Making (NDM)*, May, 2011.
- [95] Prashanth Rajivan and Nancy J Cooke. Information-pooling bias in collaborative security incident correlation analysis. *Human factors*, 60(5):626–639, 2018.

- [96] Prashanth Rajivan, Marco A Janssen, and Nancy J Cooke. Agent-based model of a cyber security defense analyst team. In *Proceedings of the human factors and ergonomics society annual meeting*, volume 57, pages 314–318. SAGE Publications Sage CA: Los Angeles, CA, 2013.
- [97] Akshitha Raman and S Ravi. A survey on intent detection methods in natural language processing. *International Journal of Advanced Computer Science and Applications*, 11(3):122–128, 2020.
- [98] Glenn D Reeder and William F Brewer. Memory for gist and exact wording in scripts: Implications for eyewitness testimony. *Journal of Experimental Psychology: Applied*, 7(4):314–327, 2001.
- [99] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 3982–3992, 2019.
- [100] F.E. Ritter, F. Tehranchi, and J.D. Oury. Act-r: A cognitive architecture for modeling cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, 10(3):1488, 2019.
- [101] Frank E Ritter, Nigel R Shadbolt, Dave Elliman, and Richard M Young. *Cognitive architectures: A way forward for modeling human cognition*, volume 3, pages 89–104. Wiley Online Library, 2011.
- [102] H. L. Roediger and A. C. Butler. The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, 15(1):20–27, 2011.
- [103] Jacqueline Strunk Sachs. Recognition memory for syntactic and semantic aspects of connected discourse. *Perception & Psychophysics*, 2(9):437–442, 1967.
- [104] Riccardo Sacripante, Robert H Logie, Alan Baddeley, and Sergio Della Sala. Forgetting rates of gist and peripheral episodic details in prose recall. *Memory & Cognition*, 51(1):71–86, 2023.
- [105] Mari Sappeli, Mikko Laitinen, Tanja Suomalainen, Sampo Kumpulainen, and Kalervo Järvelin. Assessing e-mail intent and tasks in e-mail messages. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1105–1108. ACM, 2013.
- [106] Mijke Sappeli, Hans van der Laan, Jaime Teevan, and Amy Karlson. Assessing e-mail intent and tasks in e-mail messages. *Information Systems*, 60:207–222, 2016.
- [107] B.D. Sawyer and P.A. Hancock. Hacking the human: The prevalence paradox in cybersecurity. *Human Factors*, 60(5):597–609, 2018.

- [108] Ben D Sawyer, Victor S Finomore, Greg J Funke, Vincent F Mancuso, Brent Miller, Joel Warm, and PA Hancock. Evaluating cybersecurity vulnerabilities with the email testbed: Effects of training. In *Proceedings 19th Triennial Congress of the IEA*, volume 9, page 14, 2015.
- [109] Bruce Schneier. *Data and Goliath: The Hidden Battles to Collect Your Data and Control Your World*. W. W. Norton & Company, 2015.
- [110] Tal Schuster, Ori Ram, and Regina Barzilay. Towards scalable and reliable intent classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, page 2385–2395, 2019.
- [111] Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.
- [112] Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.
- [113] S. Sheng, M. B. Holbrook, and D. Cheung. The impact of heuristics on phishing vulnerability. *Journal of Computer Information Systems*, 50(3):55–66, 2010.
- [114] Matthew Shonman, Xiaoyu Shi, Mingqing Kang, Zuo Wang, Xiangyang Li, and Anton Dahbura. Using a computational cognitive model to understand phishing classification decisions. In *35th International BCS Human-Computer Interaction Conference*, pages 1–10, 2022.
- [115] Kai Shu, Subhabrata Mukherjee, Guoqing Zheng, Ahmed Hassan Awadallah, Milad Shokouhi, and Susan Dumais. Learning with weak supervision for email intent detection. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1051–1060, 2020.
- [116] K. Singh, P. Aggarwal, P. Rajivan, and C. Gonzalez. What makes phishing emails hard for humans to detect? In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 64, pages 431–435. SAGE Publications, Los Angeles, CA, 2020.
- [117] Karan Singh, Pooja Aggarwal, Prashanth Rajivan, and Cleotilde Gonzalez. Training to detect phishing emails: Effects of the frequency of experienced phishing emails. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 63, pages 453–457. SAGE Publications Sage CA: Los Angeles, CA, 2019.
- [118] G. Sperling. The information available in brief visual presentations. *Psychological Monographs*, 74(11):1–29, 1960.

- [119] L. R. Squire. Memory and brain systems: 1969–2009. *Journal of Neuroscience*, 29:12711–12716, 2009.
- [120] Ouyang L. Wu J. Ziegler D. M. Lowe R. Voss C. Radford A. Amodei D. Stiennon, N. and P. F. Christiano. Learning to summarize from human feedback. *Advances in Neural Information Processing Systems*, 2022.
- [121] Andrea Stocco, Konstantinos Mitsopoulos, Yuxue C Yang, Holly S Hake, Theodros Haile, Bridget Leonard, and Kevin Gluck. Fitting, evaluating, and comparing cognitive architecture models using likelihood: A primer with examples in act-r. *arXiv preprint arXiv:2410.18055*, 2024.
- [122] Niels Taatgen. The atomic components of thought: by john r. anderson and christian lebiere. *Trends in Cognitive Sciences*, 3(2):82, 1999.
- [123] Robert Thomson, Christian Lebiere, John R Anderson, and James Staszewski. A general instance-based learning framework for studying intuitive decision-making in a cognitive architecture. *Journal of Applied Research in Memory and Cognition*, 4(3):180–190, 2015.
- [124] Perry W. Thorndyke and Gordon H. Bower. Storage and retrieval processes in sentence memory. *Cognitive Psychology*, 6(4):515–543, 1974.
- [125] E. Tulving. Episodic and semantic memory. In *Organization of Memory*, pages 381–403. Academic Press, 1972.
- [126] E. Tulving and D. M. Thomson. Encoding specificity and retrieval processes in episodic memory. *Psychological Review*, 80(5):352–373, 1973.
- [127] Gokhan Tur and Renato De Mori. *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech*. John Wiley Sons, 2011.
- [128] Elochukwu Ukwandu, Mohamed Amine Ben Farah, Hanan Hindy, David Brosset, Dimitris Kavallieros, Robert Atkinson, Christos Tachtatzis, Miroslav Bures, Ivan Andonovic, and Xavier Bellekens. A review of cyber-ranges and test-beds: Current and future trends. *Sensors*, 20(24):7148, 2020.
- [129] Amber Van Der Heijden and Luca Allodi. Cognitive triaging of phishing attacks. In *28th USENIX Security Symposium (USENIX Security 19)*, pages 1309–1326, 2019.
- [130] V. Vapnik and A. Vashist. A new learning paradigm: Learning using privileged information. *Neural Networks*, 22(5):544–557, 2009.

- [131] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [132] Vladislav D Veksler, Norbou Buchler, Blaine E Hoffman, Daniel N Cassenti, Char Sample, and Shridat Sugrim. Simulations in cyber-security: a review of cognitive modeling of network attackers, defenders, and users. *Frontiers in psychology*, 9:691, 2018.
- [133] Verizon. 2022 data breach investigations report, 2022.
- [134] Arun Vishwanath, Brynne Harrison, and Yu Jie Ng. Suspicion, cognition, and automaticity model of phishing susceptibility. *Communication research*, 45(8):1146–1166, 2018.
- [135] Arun Vishwanath, Tejaswini Herath, Rui Chen, Jingguo Wang, and H Raghav Rao. Why do people get phished? testing individual differences in phishing vulnerability within an integrated, information processing model. *Decision Support Systems*, 51(3):576–586, 2011.
- [136] Aikaterini Voudouri, Bence Bago, Grégoire Borst, and Wim De Neys. Reasoning and cognitive control, fast and slow. *Judgment and Decision Making*, 18:e33, 2023.
- [137] Alex Wang and Chiou Lee. Human decision-making in the age of information overload: A review of cognitive and behavioral mechanisms. *Journal of Cybersecurity and Privacy*, 3(1):10–25, 2020.
- [138] Ping Wang and Peyton Lutchkus. Psychological tactics of phishing emails. *Issues in Information Systems*, 2023.
- [139] Wei Wang, Yaliang Zhang, Ying Shen, Shuguang Ma, and Wei Li. Context-aware intent identification in email conversations. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)*, pages 2355–2363, 2019.
- [140] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in neural information processing systems*, 33:5776–5788, 2020.
- [141] Catherine Williams, Deborah Tatar, and Robert W Reeder. Cues to deception in phishing emails: The role of urgency and intent. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2019.

- [142] Peng Xu and Ruhi Sarikaya. Bert for joint intent classification and slot filling. In *Proceedings of the 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, page 716–721, 2019.
- [143] Tianhao Xu, Kuldeep Singh, and Prashanth Rajivan. 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*, volume 65, pages 1500–1504. SAGE Publications Sage CA: Los Angeles, CA, 2021.
- [144] Tianhao Xu, Kuldeep Singh, and Prashanth Rajivan. Modeling phishing decision using instance based learning and natural language processing. In *HICSS*, pages 1–10, 2022.
- [145] Tianhao Xu, Kuldeep Singh, and Prashanth Rajivan. Personalized persuasion: Quantifying susceptibility to information exploitation in spear-phishing attacks. *Applied Ergonomics*, 108:103908, 2023.
- [146] Muhammad Mudassar Yamin, Basel Katt, and Vasileios Gkioulos. Cyber ranges and security testbeds: Scenarios, functions, tools and architecture. *Computers & Security*, 88:101636, 2020.
- [147] Yuxue C Yang, Catherine Sibert, and Andrea Stocco. Reliance on episodic vs. procedural systems in decision-making depends on individual differences in their relative neural efficiency. *Computational Brain & Behavior*, 7(3):420–436, 2024.
- [148] J. Zhang, X. Zhao, and L. Wang. Cognitive biases in phishing vulnerability: A comprehensive study. *Journal of Cyberpsychology*, 6(3):225–240, 2022.
- [149] Shuai Zhang, Lina Yao, Aixin Sun, and Yi Tay. Deep learning based recommender system: A survey and new perspectives. *ACM computing surveys (CSUR)*, 52(1):1–38, 2019.
- [150] Rachael Ziegler and David Gillespie. Message interpretation and intent recognition: Effects of message tone and context on cognition. *Journal of Communication Psychology*, 44(3):219–236, 2012.

Appendix A

SURVEY QUESTIONNAIRE ITEMS

Screening and Agreement Questions

- **Subject's Statement:**

- I am 18 years or older: Yes / No

- **Read Consent:**

- I have read and understand the information above: Yes / No

- **Willing to Participate:**

- I want to participate in this research and continue with this activity: Yes / No

Introduction Shown to Participants

This study focuses on evaluating email summaries and collecting your perceptions of tone, intent, and trustworthiness.

What to Expect: You will be shown 10 emails, one at a time. After each email, you'll answer a few questions about it. You will also rate how well two summaries of the email reflect your recall of the email content.

You will also be asked to recall keywords or phrases of the email you read. The email will not be shown to you while answering this question. Please try your best to recall from memory. Please do not record, copy, or photograph the contents of the email.

Your responses will contribute to improving email management systems and understanding how people interpret email communications.

Survey Questions per Email

1. Keyword Recall Question:

Imagine a few days have passed since you read this email. What keywords would you use to find this email in your inbox?

Keywords could be any details that you can recall or remember about this email from your memory. It could be the topic being discussed in the email, specific details in the message, or information about the sender. Any key piece of information you can remember about the email.

Example: If the email was about a delayed package delivery, some possible keywords you may recall about it are: "package," "delivery," "delayed," "tracking," and "expected date." Please recall from memory when answering this question.

Please provide 5 such keywords for this email, separated by commas (e.g., package, delivery, delayed, tracking, expected date).

2. Tone Rating:

How would you rate the tone of this email on a scale from casual to formal?

Options:

- Casual
- Neutral
- Formal

3. Sentiment Rating:

What is the overall sentiment of this email?

Options:

- Negative
- Neutral

- Positive

4. Intent Selection:

What do you think was the sender’s intent with this email? (Select all the intents that you can recall.)

- Request – Asking the recipient to take an action and/or provide information (e.g., password reset, answer a question)
- Propose – Suggesting a joint action or agreement (e.g., proposing a meeting or collaboration)
- Commit – Promising to take a future action (e.g., “I’ll send it tomorrow”)
- Deliver – Providing information, recommendation, updates, or files (e.g., login alert, invoice, subscription confirmation)
- Amend – Modifying a previous proposal or decision or plan (e.g., changing meeting time or correcting info)
- Refuse – Declining a request or offer (e.g., “I can’t attend,” “Sorry, not possible”)
- Greet – Sending a social message (e.g., thank you, congratulations, greetings)
- Remind – Sending a reminder about something (e.g., upcoming deadline, follow-up)
- Resolve – Closing an issue, conflict, or question (e.g., confirming a problem is fixed, clarifying a misunderstanding, providing a final decision)

5. Summary Evaluation Questions:

This is one version of the summary of this email. Read it carefully and answer the question below.

On a scale of 1 to 5, how well does this version of the email summary match how you recall the email?

Scale: 1 (Not well at all) – 5 (Extremely well)

This is another version of the summary of this email. Read it carefully and answer the question below.

Evaluate this on its own and do not compare with the earlier summary shown to you.

On a scale of 1 to 5, how well does this version of the email summary match how you recall the email?

Scale: 1 (Not well at all) – 5 (Extremely well)