ON DESIGNING COMPUTATIONALLY ENHANCED RISK AND CRISIS COMMUNICATIONS FOR INSIDER THREATS

by

Madison Haleigh Munro

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

of

Master of Science

in

Computer Science

MONTANA STATE UNIVERSITY
Bozeman, Montana

December 2025

ACKNOWLEDGEMENTS

I would like to thank my advisor and committee chair, Dr. Ann Marie Reinhold, for her support, guidance, and patience throughout my Master's program. I would also like to thank fellow committee members Drs. Clemente Izurieta and Eric Raile and former committee member Dr. Elizabeth Shanahan, Savanna Washburn from the MSU Political Science department, all current and past members of the MSU Software Engineering and Cybersecurity Lab (SECL), research associates at the Virginia Modeling, Simulation and Analysis Center (VMASC), and Dr. Manuel Ruiz Aravena from Mississippi State University for providing support towards my research throughout the program. I also extend my gratitude to the MSU Horn Ensemble for providing me well-needed downtime from my research. Lastly, I want to extend my gratitude to my family, friends, and roommate, Sam, for their support and encouragement throughout my Master's program.

TABLE OF CONTENTS

TABLE OF CONTENTS – CONTINUED

TABLE OF CONTENTS – CONTINUED

LIST OF TABLES

LIST OF FIGURES

Figure Page

ABSTRACT

Insider threats pose significant harm to organizations of all types. The financial costs for mitigating current and future insider threats increases in the millions of dollars each year. Furthermore, insider threat mitigation strategies often target malicious insiders despite the prevalence of inadvertent insider threats. To reduce costs while also targeting inadvertent insiders, organizations can develop and deploy risk and crisis communication (RCC) to better prepare employees against insider threats.

RCC messages need to be developed and deployed effectively and efficiently. Effective messaging influences individuals to take protective actions against insider threats; efficient messaging involves swift message construction and delivery to impacted populations. While current RCC research specifies how to improve message efficacy and efficiency, much of RCC message development relies on time-consuming, laborious processes. These processes can be improved through the integration of computational text analysis and linguistic tools.

I present a methodology on designing and developing computationally enhanced RCC for insider threats. I first conducted a systematic literature review (SLR) to discover any current use of computational tools to improve RCC message efficacy and development efficiency. Next, I performed content analysis on insider threat source text using a mixed methods approach. For this approach, I leveraged Natural Language Processing (NLP) techniques and tools—including the Large Language Model (LLM) *ChatGPT*—to process text using the Narrative Policy Framework (NPF) as the guiding theoretical framework. Lastly, I constructed insider threat risk messages using the content analysis results as message content and structure. These messages were constructed using a customized version of the LLM *Llama* specialized for RCC message construction. For both content analysis and message construction steps, I evaluated how well computational tools perform message development steps. Based on my evaluations, I determined the extent to which computational tools can replace humans in RCC message development, finding that the combination of human validation and computational analysis can lead efficient development of effective messaging. By creating effective RCC messages efficiently, impacted populations can swiftly take action against organizational insider threats.

INTRODUCTION

Imagine that you are an accountant for an organization. This organization relies on you managing and safely storing spreadsheets containing quarterly financial data. On any normal day, you perform these tasks well. However, one day you slip up—not realizing until it's too late. On this day, you forget to lock your laptop before leaving for a lunch break. This critical error allows another employee with ill-intent to strike. Unbeknownst to you and your organization, this employee inserts a USB drive, copies each of your financial spreadsheets onto the drive, and leaves without detection. Days later, your compromised spreadsheets provide critical trade secrets to a competing organization, leading your organization towards financial hardship. *You have fallen victim to an insider threat.*

Insider threats refer to cybersecurity threats where an entity compromises an organization's data and IT/cybersecurity systems [29]. These threats can either be intentional or unintentional [29]. Expressions of intentional insider threats[1] include commercial and financial espionage, intellectual and financial theft, social engineering attacks, and infrastructural sabotage[2]. For unintentional insider threats[3], common incidents include falling for social engineering attacks, improper handling of data, accidentally disclosing sensitive organizational information, and misplacing physical data storage [30]. Insider threats, intentional or not, can pose significant harm to organizations of all types—from corporate to governmental[4].

Organizational insider threats present significant financial and infrastructural cost. In

[1] Also known as malicious insider threats

[2] https://www.cisa.gov/resources-tools/resources/insider-threat-mitigation-guide

[3] Also known as inadvertent insider threats

[4] https://gurucul.com/blog/what-is-an-insider-threat/

2023, the average financial cost of insider threats totaled around 16.2 *million* dollars[5]; this cost has since risen to around 17.4 *million* dollars in 2025[6]. Regarding infrastructural cost, insider threats have made organizations vulnerable, sometimes damaging their reputation. As a recent example, in 2023, an insider compromised the U.S. Department of Defense (DoD) by leaking sensitive and classified military data, putting the country in serious jeopardy from foreign adversaries[7]. To prevent future harm caused by insider threats, many organizations implement threat mitigation tools and systems.

Myriad techniques are adopted by organizations to counter or prevent insider threats. Such countering and preventative techniques rely on technical measures such as computational anomaly or signature threat detection [5, 79], monitoring data logs and network traffic [5, 29, 53], or statistical analysis (e.g., time series analysis, clustering) [5, 53]. Many countering techniques also take into account behavioral factors of insider threats such as employee behavior analysis and checking organizational policies and employee records [5, 29, 30]. However, organizations often deploy insider threat detection and mitigation systems in response to active insider threat attacks, primarily targeting malicious insiders despite the frequency of inadvertent insider incidents within organizations[5,6]. Furthermore, deploying such systems can cost organizations millions of dollars[5,6], compounding any financial losses from initial insider attacks. To better target inadvertent insider threats while reducing financial costs, organizations can develop and deploy risk communication messages to prepare employees *before* insider threats strike.

Risk and crisis communication (RCC) message development spans across many hazard domains—including general cybersecurity [36, 81, 114] and insider threats [68, 85]. Effective RCC messaging encourages impacted populations to take preventative or protective action when faced with a hazard [48, 67, 83, 95]. In tandem with message efficacy is efficient message

[5] https://www.syteca.com/en/blog/insider-threat-statistics-facts-and-figures
[6] https://www.dtexsystems.com/blog/2025-cost-insider-risks-takeaways/
[7] https://edition.cnn.com/2023/04/13/politics/us-government-intel-leak/index.html

development, delivery, and deployment [54, 82]. While calls for improving message efficacy are common in RCC research [67], most RCC research still relies on manual, inefficient development approaches [83]. Manual development of RCC messages typically present issues with study validity as studies rely on black-box approaches for message construction and analysis [83]. This approach towards development is also vague about *how* messages are developed, resulting in messages being constructed with unmitigated threats to precision and quality [82, 83]. Nevertheless, newer RCC research have taken steps towards efficiently developing precise and quality messaging through the integration of Natural Language Processing (NLP) and theoretical frameworks grounded in social science [19, 43, 44, 56–59, 67, 68, 71, 82, 83, 95]. One potential solution towards efficient development of effective messaging is the Domain Agnostic Risk Communication (DARC) Framework [82].

The DARC Framework enables the continued collaboration of NLP and theoretical frameworks to analyze and construct RCC messages while promoting instrument fidelity [82]. Furthermore, the DARC Framework adheres to software engineering principles such as encapsulation, abstraction, and extensibility so that RCC researchers can apply the framework to any hazard, framework, and computational text analysis tools of choice [82]. However, to date, practical applications of the DARC that test its applicability and validity in RCC development steps are nascent. To that end, I present a practical application of the DARC Framework with insider threats as the hazard of choice within this thesis.

This thesis is driven by three primary objectives: (1) identify current and past uses of computational tools in RCC message development and present the state of the science, (2) perform content analysis on insider threat source text using computational tools and theoretical frameworks widely used by RCC researchers, and (3) construct RCC messages on insider threats efficiently using computational tools to promote instrument fidelity. To better organize my research objectives, I developed a hybrid model based on Basili's Goal-Question-Metric (GQM) approach [10] and Schimel's Question-Challenges-Objectives outline

[90] (Figure 2.1) called Goal-Question-Objectives (GQO). My research GQO address four research questions, each answered in the manuscripts presented in this thesis.

The first manuscript, “Enhancing Risk and Crisis Communication with Computational Methods: A Systematic Literature Review”, presents a systematic literature review (SLR) identifying current and past uses of computational tools during any stage of RCC development. Additionally, the SLR identifies current and past applications of theoretical frameworks used to guide RCC development. Both findings in the SLR inform research questions 1 and 2 with answers provided by objectives 1 and 2, respectively (Figure 2.1).

The second manuscript, “Integrating Computational Text Analysis in Risk and Crisis Communication Development”, presents the integration of computational tools and a selected RCC message framework into RCC content analysis steps. The computational tools include various NLP techniques, the Large Language Model (LLM) *ChatGPT*[8], and statistical modeling and analysis tools; the chosen RCC message framework is the Narrative Policy Framework (NPF). With this integration, I evaluate the extent to which computational tools replicate human-based (manual) analysis, determining if such tools are capable of replacing humans. This work informs research question 3, which is answered by objectives 3 and 4 (Figure 2.1).

The last manuscript, “Persuasion with More Precision: Leveraging Large Language Models to Improve Insider Threat Risk Communication”, evaluates the efficiency of risk message construction using the LLM *Llama*[9] while promoting instrument fidelity. The content and structure of these messages is derived from the content analysis results presented in the second manuscript, and the LLM used is a custom instance of *Llama* specialized for insider threat risk communication. This work informs research question 4, which is answered by objectives 5 and 6 (Figure 2.1).

[8] https://openai.com/chatgpt/overview/

[9] https://www.llama.com/models/llama-3/

The GQO, manuscript #1, manuscript #2, and manuscript #3 are presented in the following sections along with concluding remarks on RCC for insider threats.

RESEARCH GOALS

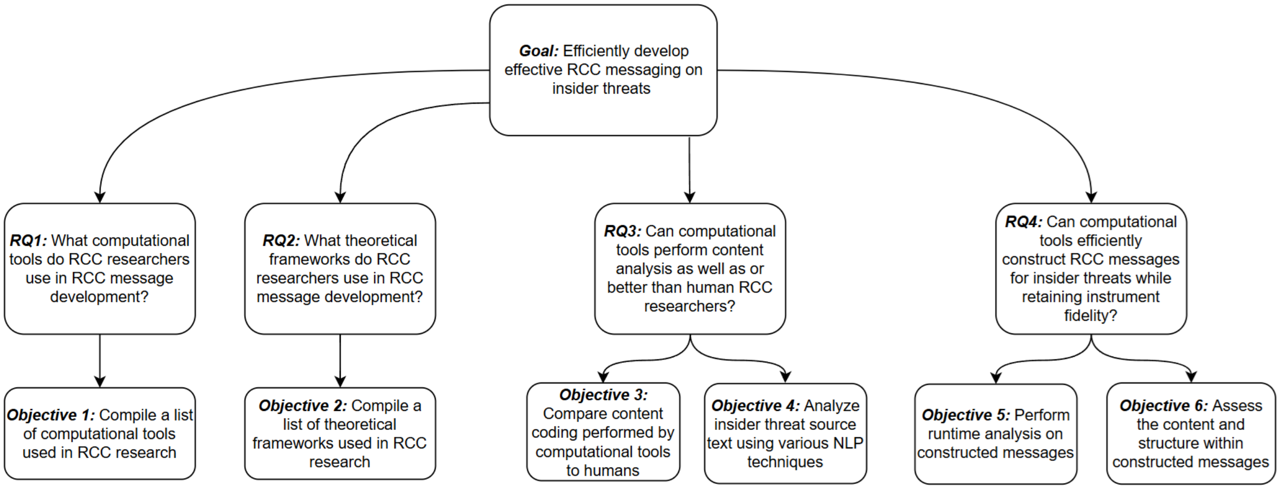
Goal: Efficiently develop effective RCC messaging on insider threats
RQ1: What computational tools do RCC researchers use in RCC message development?
RQ2: What theoretical frameworks do RCC researchers use in RCC message development?
RQ3: Can computational tools perform content analysis as well as or better than human RCC researchers?
RQ4: Can computational tools efficiently construct RCC messages for insider threats while retaining instrument fidelity?
Objective 1: Compile a list of computational tools used in RCC research
Objective 2: Compile a list of theoretical frameworks used in RCC research
Objective 3: Compare content coding performed by computational tools to humans
Objective 4: Analyze insider threat source text using various NLP techniques
Objective 5: Perform runtime analysis on constructed messages
Objective 6: Assess the content and structure within constructed messages

Figure 2.1: Hierarchical Goal-Question-Objectives (GQO) structure based on Basili's Goal-Question-Metric and Schimel's Question-Challenges-Objectives methodologies [10, 90].

To ensure a well-structured research plan, I employed a hybrid methodology based on both Basili's Goal-Question-Metric approach [10] and Schimel's Question-Challenges-Objectives outline [90]. This hybrid model defines my overarching research goal: efficiently develop effective RCC messaging on insider threats. To address this goal, I created four research questions, each answered by one or more objectives. Research questions 1 and 2 corresponded to the manuscript in Chapter 3, research question 3 corresponded to the manuscript in Chapter 4, and research question 4 corresponded to the manuscript in Chapter 5. The structure of this plan is detailed in Figure 2.1 and in the following text:

Goal: Efficiently develop effective RCC messaging on insider threats.

- **Research Question 1:** What computational tools do RCC researchers use in RCC message development?
  - **Objective 1:** Compile a list of computational tools used in RCC research.

- **Research Question 2:** What theoretical frameworks do RCC researchers use in RCC message development?
  - **Objective 2:** Compile a list of theoretical frameworks used in RCC research.
- **Research Question 3:** Can computational tools perform content analysis as well as or better than human RCC researchers
  - **Objective 3:** Compare content coding performed by computational tools to humans.
  - **Objective 4:** Analyze insider threat source text using various NLP techniques.
- **Research Question 4:** Can computational tools efficiently construct RCC messages for insider threats while retaining instrument fidelity?
  - **Objective 5:** Perform runtime analysis on constructed messages.
  - **Objective 6:** Assess the content and structure within constructed messages.

ENHANCING RISK AND CRISIS COMMUNICATION WITH COMPUTATIONAL METHODS: A SYSTEMATIC LITERATURE REVIEW

Contribution of Authors and Co-Authors

Manuscript in following chapter

Author: Madison H. Munro

Contributions: Developed study concept and design, data collection and analysis, interpretation of results, and wrote the manuscript.

Co-Author: Ross J. Gore

Contributions: Assisted with the literature review process, provided feedback, and assisted with editing.

Co-Author: Christopher J. Lynch

Contributions: Assisted with the literature review process, provided feedback, and assisted with editing.

Co-Author: Yvette D. Hastings

Contributions: Assisted with the literature review process, provided feedback, and assisted with editing.

Co-Author: Ann Marie Reinhold

Contributions: Provided general study direction, feedback, and substantial edits on the manuscript.

Manuscript Information

Madison H. Munro, Ross J. Gore, Christopher J. Lynch, Yvette D. Hastings,

Ann Marie Reinhold

Publisher: Risk Analysis - An International Journal
In-press
Originally published in December 2024; officially published in July 2025
Volume 45, Issue 7
DOI: doi/10.1111/risa.17690

Status of Manuscript:
____ Prepared for submission to a peer-reviewed journal
____ Officially submitted to a peer-reviewed journal
____ Accepted by a peer-reviewed journal
_X_ Published in a peer-reviewed journal

Abstract

Recent developments in risk and crisis communication (RCC) research combine social science theory and data science tools to construct effective risk messages efficiently. However, current systematic literature reviews (SLRs) on RCC primarily focus on computationally assessing message efficacy as opposed to message efficiency. We conduct an SLR to highlight any current computational methods that improve message construction efficacy and efficiency. We found that most RCC research focuses on using theoretical frameworks and computational methods to analyze or classify message elements that improve efficacy. For improving message efficiency, computational and manual methods are only used in message classification. Specifying the computational methods used in message construction is sparse. We recommend that future RCC research apply computational methods toward improving efficacy and efficiency in message construction. By improving message construction efficacy and efficiency, RCC messaging would quickly warn and better inform affected communities impacted by current hazards. Such messaging has the potential to save as many lives as possible.

Introduction

Risk and crisis communication (RCC) is a powerful tool for improving hazard preparedness. RCC encourages individuals to take protective actions to keep themselves and their community safe [23, 39, 82, 84]. Such communication encompasses both risk messaging deployed before a hazard and crisis messaging deployed during a disaster. *Effective* RCCs motivate as many individuals as possible in a target population to adopt protective actions.

Efficacy (definition in Table 3.1) is essential in RCC. Achieving messaging efficacy involves bridging knowledge gaps between hazard domain experts and affected populations [83]. Bridging such gaps relies on how the message is developed. This development frames RCC objectives in ways such that affected populations are receptive to and engaged with the messaging. Receptivity and engagement motivate affected populations to take action to protect themselves from a hazard [37, 94].

Table 3.1: Definitions for key terms used in this systematic literature review (SLR), as defined by the authors and other sources.

Term	**Definition**	**Definitions inferred from other sources**
Efficacy	The extent to which risk messaging changes individuals' risk perceptions and mitigation behavior when faced with a hazard	Measuring persuasive outcomes of risk messages, including changes in attitudes, behaviors, intentions, and knowledge of individuals [108]; Risk communication that incorporates honesty, reassurance, and actionable items in its messaging, usually guided by a risk communication framework [1].
Efficiency	Combines the speed with which messages can be created with optimal use of resources	A general set of standards, procedures, guidelines, norms, reference points, or principles that are designed to improve performance [91]; The ability for a computational method to complete its function that is both computationally inexpensive and scalable for large-scale data [50]; Improvements to work or task performance through automation and aggregation of tasks [109].

RCC also relies on the timely delivery of messages [54]. Messages disseminated promptly ensure individuals have as much lead time as possible to prepare for a risk or to respond during a crisis. For this reason, the time it takes to develop a message is an important concern. *Efficient* RCCs are created with optimal use of computational resources and time.

Efficiency (definition in Table 3.1) is a crucial aspect of RCCs. Achieving efficiency involves constructing RCC messages using semi-automated or fully automated computational methods [43]. Such methods enable shorter message development timeframes compared to manually constructing messages, thus resulting in timelier distribution of messages to populations impacted by hazards. Successful RCC depends on how researchers create effective and efficient messaging.

Recent developments in RCC research blend the worlds of psychology, policy process, and data science to construct effective risk messages efficiently [27, 82]. Although there is growing emphasis on improving message efficacy using computational methods [70, 83], RCC research prioritizes social science aspects informing effective message development. This prioritization is specifically on analysis of effective message elements [9, 21, 34].

Computational methods for analyzing message efficacy are well researched [32, 72], but effective computational message construction is largely ignored [83]. This is not to say that message efficacy has been overlooked, but that the use of computational methods improving message efficacy is limited. In addition, little research attention has been paid to efficient message construction—both in using computational or manual methods. One reason for this gap in efficiency research is because computational methods have recently become prominent in the last few years [41]. Overall, research into computational message construction is sparse, a problem that is reflected in systematic literature reviews (SLRs) on RCC research [9, 21, 34, 72].

Existing SLRs aggregate research on message efficacy, specifically on manual [9, 21, 34] or computational [72] classification and analysis of message elements. However, none of them focus on message construction. To date, no SLRs investigate methods used in message construction for improving efficacy, nor do any SLRs investigate methods for improving the efficiency of message construction.

To address these critical research gaps, we conduct an SLR to investigate what, if any, research focuses on efficacy and efficiency in message construction. We highlight existing gaps present in research on computational message construction. Further, we highlight usages of computational methods in RCC research. The next section details the methodology this SLR coincides with.

Methods

The methodology for this SLR aligned with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses[1] (PRISMA) guidelines [74, 75]. We primarily adhered to guideline steps aligning with data synthesis. These steps include developing research questions (RQs), developing a search strategy, assessing eligibility, performing data meta-analysis, and selecting our final literature to report.

Guideline Selection and Research Questions

The first step in PRISMA facilitated the development of RQs on computational methods used in message construction. The following RQs drove the direction of this SLR:

RQ1. What established and emerging computational methods improve efficacy in RCC message construction?

RQ2. What established and emerging computational methods improve efficiency in RCC message construction?

[1] http://prisma-statement.org/

RQ3. What are the trade-offs between efficient and effective computational methods for message construction?

All three RQs reflect our research attention on effective and efficient computational methods used in RCC. After RQ development, we focused on the next step in the PRISMA guidelines: developing a search strategy.

Selection, Search, and Screening

Database Selection and Search We constructed search strings from both Boolean logical operators and terms derived from the defined RQs. Terms within the strings cover RCC, messaging, and computational methods. Search string formatting varied across selected databases (Table 3.2).

We utilized the research databases *IEEE Xplore*[2], *ACM Digital Library*[3], *Web of Science*[4], and *PubMed*[5] to find relevant RCC message literature. *IEEE Xplore* and *ACM Digital Library* were chosen for their collection of multidisciplinary computer science research; *Web of Science* and *PubMed* were chosen for their collection of multidisciplinary social science research. The database searches uncovered 1,385 potentially relevant literature on message construction (548 from *IEEE Xplore*, 299 from *ACM Digital Library*, 313 from *Web of Science*, and 225 from *PubMed*).

Results Screening Literature identified from each database underwent automatic and manual screening. Filters for publishing year and publication type narrowed the number of articles down to 640. Filtered articles had their titles and DOI links web scraped into dataframes corresponding to the database the article was identified from. The scraping was

[2] https://ieeexplore.ieee.org/Xplore/home.jsp
[3] https://dl.acm.org/
[4] https://www.webofscience.com/wos/woscc/basic-search
[5] https://pubmed.ncbi.nlm.nih.gov/

Table 3.2: Searched databases and respective strings used to query results

Database used	Search string	Filters applied	Results	Date searched
ACM Digital Library	[[All: “risk communication”] OR [All: “crisis communication”] OR [All: “hazard communication”]] AND [All: messag] AND [[All: analysis] OR [All: computational analysis] OR [All: “natural language processing”] OR [All: “nlp”] OR [All: “artificial intelligence”] OR [All: “ai”]] AND [E-Publication Date: (01/01/2018 TO 12/31/2023)]	Research articles only	138	2/7/2024
IEEE Xplore	(“Full Text & Metadata”:“risk communication” OR “Full Text & Metadata”: “crisis communication” OR “Full Text & Metadata”:“hazard communication”) AND (“Full Text & Metadata”:messag) AND (“Full Text & Metadata”:analysis OR “Full Text & Metadata”:computational analysis OR “Full Text & Metadata”:“natural language processing” OR “Full Text & Metadata”:“nlp” OR “Full Text & Metadata”:“artificial intelligence” OR “Full Text & Metadata”:“ai”)	Published between 2018 and 2023, journals and conferences only	166	2/7/2024
PubMed	((“risk communication” OR “crisis communication” OR “hazard communication”) AND (messag)) AND (analysis OR computational analysis OR “natural language processing” OR “nlp” OR “artificial intelligence” OR “ai”)	Published between 2018 and 2023	144	2/7/2024
Web of Science	((AB=(“risk communication” OR “crisis communication” OR “hazard communication”)) AND AB=(messag)) AND AB=(analysis OR computational analysis OR “natural language processing” OR “nlp” OR “artificial intelligence” OR “ai”)	Published between 2018 and 2023, articles only	192	2/7/2024

Note: The table also mentions any filters applied to results, the number of results after applying filters, and the search date for each database consulted. The search strings differed from each other since each database utilized different query formats. Any *'s present in search strings indicated that the term preceding it was stemmed.

done both manually and using the *R*[6] package *rvest*[7]. *R* was also used for postprocessing to filter out duplicate and invalid entries in each dataframe, reducing the number of articles from 640 to 627. Articles selected after the screening were subsetted and dispersed to six reviewers recruited for manual evaluation.

Manual Eligibility Analysis

Inclusion Criteria We developed inclusion criteria to assess the relevance of the screened literature (Table 3.3). The criteria specified which topics and characteristics research-relevant literature needed to include. For example, a relevant literature source needed to discuss computational RCC messaging, have an English publication, be published between January 2018 and December 2023, and be a primary literature source. Such characteristics were chosen to ensure that the most recent, state-of-the-art, and high-impact literature was captured. The criteria assisted reviewers with assessing article relevance based on their abstracts.

Abstract Reviews Disbursement of literature to reviewers occurred as follows. For this study, the lead author reviewed 317 abstracts, 2 co-authors reviewed 100, and 1 co-author reviewed 70. Two additional reviewers each read 20 abstracts to lessen the workload of the main reviewers. Three reviewers with RCC knowledge took on most of the abstract review load (at least 100 abstracts), and the remaining three were doctoral students; every reviewer had knowledge about computational sciences. All reviewers worked independently of each other with no overlap, meaning that no ties among reviewers could occur.

Prior to evaluating abstracts, all reviewers received a copy of the inclusion criteria (Table 3.3). Reviewers read each article's abstract to determine research relevance. If the abstract

[6] R Core Team (2023). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

[7] Wickham H (2024). rvest: Easily Harvest (Scrape) Web Pages. R package version 1.0.4, https://CRAN.R-project.org/package=rvest

Table 3.3: Inclusion and exclusion criteria developed for the systematic literature review (SLR).

Criteria	**Status**
Study/Research topic is on risk, crisis, or hazard communication	*Inclusion*
Article covers computational or mixed methods used in research on the above topic	*Inclusion*
Article covers computational or mixed methods used specifically in risk message construction/development	*Inclusion*
Article is published between 2018 and 2023	*Inclusion*
Article is published in English	*Inclusion*
Article is a research article from primary literature	*Inclusion*
Article is not retracted, outdated, or pre-published	*Inclusion*
Article comes from a journal or conference proceeding	*Inclusion*
Article does not meet all the above criteria	*Exclusion*

was unclear, reviewers assessed the introduction for relevance. Reviewers recorded each criterion that each article met in their spreadsheets. Criteria recorded as “YES” indicated that the article was potentially relevant; criteria recorded as “NO” indicated that the article should be excluded from the SLR. Reviewers returned their spreadsheets to the lead author after completion. The lead author did a brief quality check to ensure the spreadsheets were filled in properly, removing any duplicate entries. A total of 124 articles met the inclusion criteria and thus formed our corpus. These articles underwent meta-analysis to further assess relevance.

Meta-Analysis

We conducted the meta-analysis primarily using the *R* package *tidytext*[8]. As part of the meta-analysis, titles and hyperlinks for our corpus of articles were aggregated into one dataframe. Each entry contained full text that was manually scraped from corresponding hyperlinks. The text underwent data preprocessing and cleaning before text analysis and term tokenization. Tokenized terms classified as stopwords (e.g., “and,” “the,” and “2020”) were filtered out before calculating term frequencies.

We calculated the frequency of each term across all articles by dividing the occurrence of each term in each by the total number of terms present in the same article. Term frequency calculations were foundational to calculating term frequency-inverse document frequency[9] (TF-IDF) scores for terms. The top 50 TF-IDF scores corresponded to the terms used most in our corpus (Figure 3.1). We then manually tagged terms that were relevant to our RQs (e.g., “communication,” “computational,” and “efficacy”; herein, “research-relevant terms”). Terms with both a high TF-IDF score and relevance to our RQs determined the next selection of literature. In total, 51 articles from our corpus contained research-relevant terms and were

[8] Silge, J & Robinson, D (2024). Text Mining using ’dplyr’, ’ggplot2’, and Other Tidy Tools. R version 0.4.2, https://cran.r-project.org/web/packages/tidytext/tidytext.pdf

[9] https://www.tidytextmining.com/tfidf.html

thus selected to undergo manual quality assessment.

Meta-Analysis

We conducted the meta-analysis primarily using the *R* package *tidytext*[10]. As part of the meta-analysis, titles and hyperlinks for our corpus of articles were aggregated into one dataframe. Each entry contained full text that was manually scraped from corresponding hyperlinks. The text underwent data preprocessing and cleaning before text analysis and term tokenization. Tokenized terms classified as stopwords (e.g., “and,” “the,” and “2020”) were filtered out before calculating term frequencies.

We calculated the frequency of each term across all articles by dividing the occurrence of each term in each by the total number of terms present in the same article. Term frequency calculations were foundational to calculating term frequency-inverse document frequency[11] (TF-IDF) scores for terms. The top 50 TF-IDF scores corresponded to the terms used most in our corpus (Figure 3.1). We then manually tagged terms that were relevant to our RQs (e.g., “communication,” “computational,” and “efficacy”; herein, “research-relevant terms”). Terms with both a high TF-IDF score and relevance to our RQs determined the next selection of literature. In total, 51 articles from our corpus contained research-relevant terms and were thus selected to undergo manual quality assessment.

Article quality assessment

Selected articles underwent manual quality assessment of topical findings and credibility. Regarding topical findings, final inclusion criteria required that each article presented quantitative assessments of computational methods improving RCC message efficacy and/or efficiency. If an article did not present such quantitative assessments, the article was

[10] Silge, J & Robinson, D (2024). Text Mining using 'dplyr', 'ggplot2', and Other Tidy Tools. R version 0.4.2, https://cran.r-project.org/web/packages/tidytext/tidytext.pdf

[11] https://www.tidytextmining.com/tfidf.html

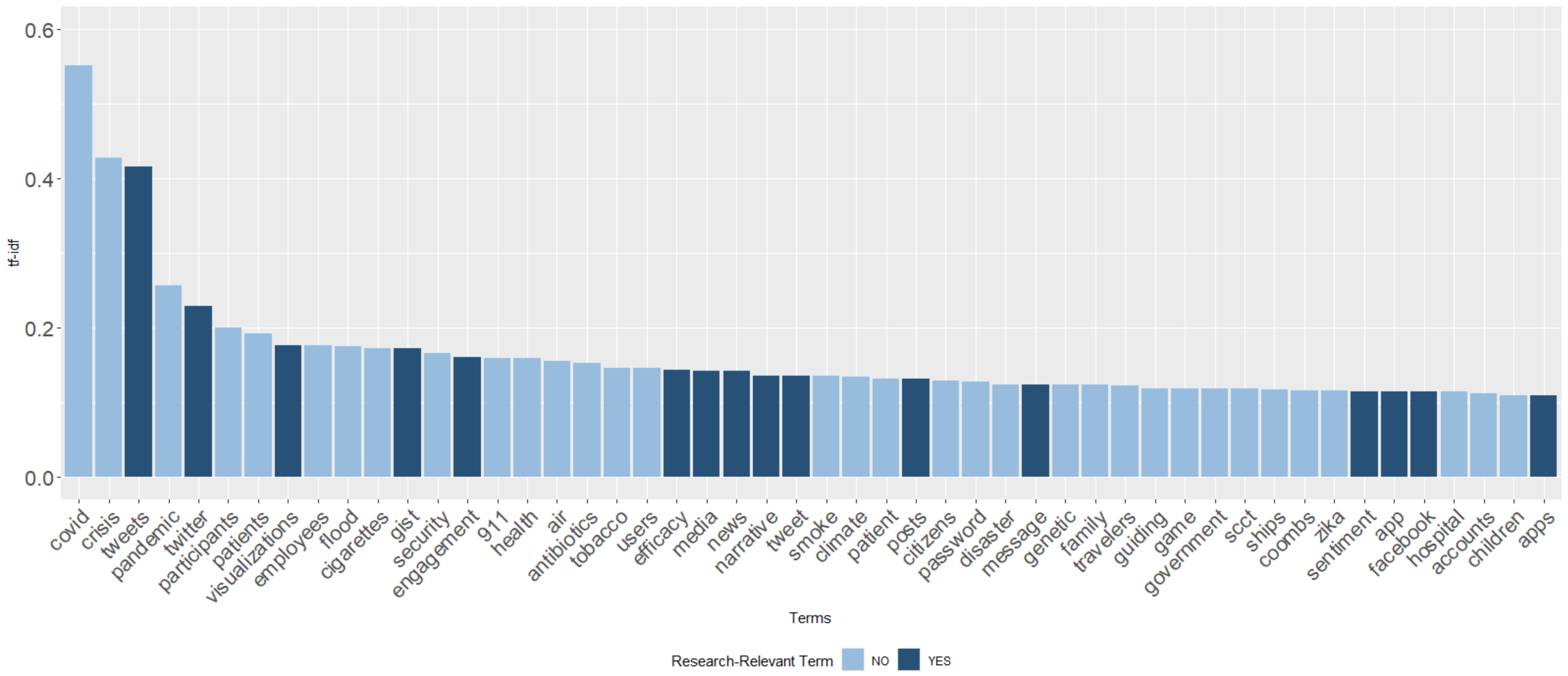

Figure 3.1: Top 50 terms based on highest term frequency-inverse document frequency (TF-IDF) scores across our corpus of articles. Research-relevant terms are shaded dark blue. Determining research relevance involved manual tagging of terms in R.

excluded. Regarding credibility, final inclusion criteria required that the presentation of the article was coherent and well-reasoned. If grammatical errors, misspellings, illogical structure, or vagaries prevented the lead author from understanding the study, the article was excluded. Based on these criteria, 25 articles were excluded from the study. Twenty-six articles made up the final selection of literature on computational risk message construction (Figure 3.2 shows how many articles were filtered out for each step in the SLR; also see Table 3.4 for article titles, authors, and topics addressed in the final selection of literature). Findings are summarized in the next section.

Table 3.4: Final selection of literature on computational risk message construction identified in this systematic literature review (SLR)

Article title	**Authors**	**Article topic**
Arabic Twitter Corpus for Crisis Response Messages Classification	[2]	Researchers developed a corpus in the Arabic language to classify crisis communication messages/tweets based on what crisis category they fell under
Comparing the Effectiveness of Text-based and Video-based Delivery in Motivating Users to Adopt a Password Manager	[3]	Researchers compared the efficacy of text- and video-based risk communication for motivating users to adopt password managers
The Saudi Ministry of Health's Twitter Communication Strategies and Public Engagement During the COVID-19 Pandemic: Content Analysis Study	[4]	Researchers aimed to evaluate the Saudi Arabia Ministry of Health's use of Twitter and the public's engagement during different stages of the COVID-19 pandemic in Saudi Arabia. They classified tweets based on the CERC framework
Hacked Time: Design and Evaluation of a Self-Efficacy Based Cybersecurity Game	[14]	Researchers developed a game that provided an interactive risk communication approach to improve risk perception of cybersecurity threats and self-efficacy in users affected
Platform Effects on Public Health Communication: A Comparative and National Study of Message Design and Audience Engagement Across Twitter and Facebook	[16]	Researchers analyzed risk communication messages dispersed by government accounts on Facebook and Twitter, specifically looking at how the public engaged with these messages
Emotionality in COVID-19 crisis communication from authorities and independent experts on Twitter	[19]	Researchers analyzed the sentiment (negative, neutral, or positive) of tweets from German health organizations during the early stages of COVID-19
Knowing Your Audience: A Typology of Smoke Sense Participants to Inform Wildfire Smoke Health Risk Communication	[35]	This study explored perspectives on wildfire smoke as a health risk among participants of Smoke Sense, a citizen science project with an objective to engage affected individuals on wildfire smoke. Researchers then developed effective health risk communication strategies to motivate individual-level behavior change

Continued on next page

Continued from previous page

Article title	**Authors**	**Article topic**
Developing a gist-extraction typology based on journalistic lead writing: A case of food risk news	[40]	This study aimed to construct a journalistic gist extraction typology to improve the development of risk communication messages. Researchers aimed to translate expert jargon into a format that was easy to read and digest for the public at large
Validation of mobile phone text messages for nicotine and tobacco risk communication among college students: A content analysis	[44]	Researchers constructed text messages for tobacco risk communication based on three main structures: framing (gain- or loss-framed messages), depth (simple or complex messages), and appeal (emotional or rational messages)
Canadian COVID-19 Crisis Communication on Twitter: Mixed Methods Research Examining Tweets from Government, Politicians, and Public Health for Crisis Communication Guiding Principles and Tweet Engagement	[57]	This study described how crisis actors used guiding principles in COVID-19 tweets and how the use of these guiding principles related to tweet engagement. Researchers classified tweets based on said guiding principles
Examining Social Media Crisis Communication during Early COVID-19 from Public Health and News Media for Quality, Content, and Corresponding Public Sentiment	[58]	Researchers aimed to evaluate the quality and content of Canadian public health and news media crisis communication during the first wave of the COVID-19 pandemic on Facebook and the subsequent emotional response to messaging by the public
A content analysis of Canadian influencer crisis messages on Instagram and the public's response during COVID-19	[59]	Researchers examined COVID-19-related crisis messages across Canadian influencer accounts on Instagram to examine their efficacy based on message constructs outlined by the Health Belief and Extended Parallel Processing models. Researchers also analyzed audience sentiment
Machine Learning Framework for Analyzing Disaster-Tweets	[62]	This study analyzed the performance of computational classifier models when classifying types of disaster crisis tweets
Build community before the storm: The National Weather Service's social media engagement	[73]	This study examined crisis communication on social media by observing how 12 National Weather Service (NWS) offices used Twitter to facilitate engagement with stakeholders during threat and nonthreat periods
Narrative Risk Communication as a Lingua Franca for Environmental Hazard Preparation	[77]	Researchers developed a new risk communication framework that guides the construction of risk messages both using narrative structure and invoking narrative transport
Investigating the presentation of uncertainty in an icon array: A randomized trial	[80]	Researchers analyzed the efficacy of visual risk communication about the risks of breast and ovarian cancer for individuals carrying the BRCA1 pathogenic variant
User-Generated Crisis Communication: Exploring Crisis Frames on Twitter during Hurricane Harvey	[87]	Researchers analyzed user-generated crisis communication—as well as crisis communication distributed by organizations—to get a well-founded understanding of how the public views risks and crises and what information was sought after
Communicating risk of medication side-effects: role of communication format on risk perception	[89]	This study assessed the interaction effects of message format and contextual factors (rate of occurrence and severity) on risk perception of medication side-effects after considering message format and contextual factors influencing risk perception

Continued on next page

Continued from previous page

Article title	**Authors**	**Article topic**
Characters matter: How narratives shape affective responses to risk communication	[95]	Researchers analyzed the use and effectiveness of narrative elements in flood risk messaging. Subsequent messages were constructed, aiming to improve individual affective response and changes in intended behavior and risk perception
A machine learning approach to flood severity classification and alerting	[97]	Researchers leveraged several machine learning models and assessed their performance at classifying flood risk message types (advisory, information, warning, and watch)
Examining Tweet Content and Engagement of Canadian Public Health Agencies and Decision Makers During COVID-19: Mixed Methods Analysis	[101]	This study examined the content and engagement of COVID-19 tweets authored by Canadian public health agencies and decision makers, making suggestions on how to improve the efficacy of crisis communication based on the results
Qualitative analysis of visual risk communication on twitter during the Covid-19 pandemic	[102]	Researchers investigated how visual risk communication was used on Twitter to promote the World Health Organization's (WHO) recommended preventative behaviors and how this communication changed over time
Story mapping and sea level rise: listening to global risks at street level	[103]	This study described the development of an interactive tool that juxtaposed coastal residents' video-recorded stories about sea level rise and coastal flooding with an interactive map that showed future sea level rise projections
An application of the extended parallel process model to protective behaviors against COVID-19 in South Korea	[111]	This study applied the EPPM to understand factors that affect an individual's participation in protective behaviors against COVID-19. Such factors included the effect of public perception of threat, the efficacy of fatalism, and undertaking protective behaviors
Sharing health risk messages on social media: Effects of fear appeal message and image promotion	[115]	This study examined how fear appeal and individuals' image promotion consideration drove users' intentions to share fear appeal messages on social networking sites
Understanding motivated publics during disasters: Examining message functions, frames, and styles of social media influentials and followers	[116]	Researchers analyzed how different message functions in risk and crisis communication were employed by Twitter users, both general users and popular influencers, using the Ariana Grande concert bombing event as the hazard subject

Note: Included are the 26 article titles, authors, and the topics discussed therein.

Abbreviations: CERC, Crisis and Emergency Risk Communication; EPPM, Extended Parallel Process Model.

Results

We report the final selection of literature identified from the SLR. Findings presented within the first three subsections provide answers to our three RQs. These RQs ask what computational methods improve message construction efficacy and efficiency as well as what

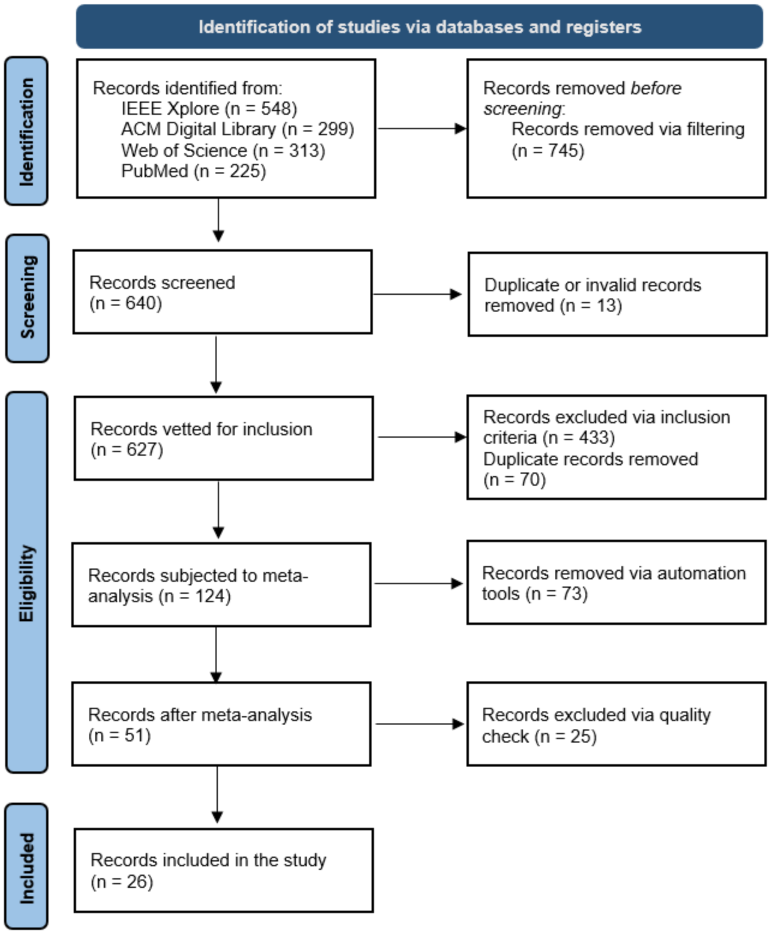
Identification of studies via databases and registers
Identification
Records identified from:
IEEE Xplore (n = 548)
ACM Digital Library (n = 299)
Web of Science (n = 313)
PubMed (n = 225)
Records removed *before screening*:
Records removed via filtering
(n = 745)
Screening
Records screened
(n = 640)
Duplicate or invalid records removed (n = 13)
Eligibility
Records vetted for inclusion
(n = 627)
Records excluded via inclusion criteria (n = 433)
Duplicate records removed
(n = 70)
Records subjected to meta-analysis (n = 124)
Records removed via automation tools (n = 73)
Records after meta-analysis
(n = 51)
Records excluded via quality check (n = 25)
Included
Records included in the study
(n = 26)

Figure 3.2: Flowchart detailing each step of the systematic literature review (SLR) process for selecting relevant literature. The figure is based on the flowchart structure specified in the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines.

trade-offs between them exist. We also present findings from the selected literature discussing other applications of computational methods in RCC.

Computational Methods and Message Construction Efficacy

Risk and crisis message construction utilizing computational methods is rarely discussed in RCC research. This SLR identified seven studies explicitly discussing computational methods used in RCC message construction [14, 44, 77, 89, 95, 103]. Furthermore, studies on computational methods used in message construction discussed only efficacy, not efficiency.

Studies on message efficacy focused on combining social science theory and computational tools to construct risk messages [14, 77, 89, 95], to analyze risk messages [3, 4, 16, 19, 40, 57–59, 73, 87, 101, 102, 116], or to analyze audience interaction with RCC messaging [3, 16, 35, 57–59, 73, 87, 101, 111, 115, 116]. Natural language processing (NLP) and content analysis were the most prevalent computational methods addressed for message construction, message analysis, and audience response analysis (Table 3.5).

Table 3.5: Effective and efficient computational methods used for risk message construction, classification, and analysis across all selected articles

Message aspect addressed	**Computational methods used**	**Articles**
Message classification efficiency	Random forest, Naïve Bayes, support vector machine, logistic regression, extreme gradient boost, decision tree	[2, 62, 97]
Message classification efficacy	Content analysis, natural language processing, random forest, Naïve Bayes, support vector machine, logistic regression, Extreme Gradient Boost, decision tree, cluster analysis, linguistic inquiry, and word count	[2–4, 16, 44, 57–59, 62, 101]
Message construction efficacy	Content analysis, natural language processing, transformational game design and programming, icon arrays, interactive story map development, linguistic inquiry, and word count	[14, 44, 77, 80, 89, 95, 103]
Message analysis efficacy	Content analysis, chi-squared analysis, natural language processing, cluster analysis, logistic regression, multiple regression, ANOVA/ANCOVA	[3, 4, 16, 19, 35, 40, 57–59, 73, 87, 101, 102, 111, 115, 116]

Note: Some articles discussed more than one aspect of message development.

Computational methods helped operationalize theoretical frameworks for effective RCC messaging. Theoretical frameworks were embedded in codebook creation and generally had two aims: (1) to improve the efficacy of RCC messaging or (2) to identify effective elements in RCC messages. These codebooks were used for message construction [14, 44, 77, 89, 95] and message analysis [3, 4, 16, 19, 35, 40, 57–59, 73, 87, 102, 111, 115, 116]. The Crisis and Emergency Risk Communication (CERC) model, the Narrative Policy Framework (NPF), the extended parallel process model (EPPM), and protection motivation theory (PMT) were the most prevalent operationalized theoretical frameworks (Table 3.6).

Table 3.6: Theoretical frameworks discussed and implemented in risk message development across all final selections of articles

Theoretical framework	**Article(s)**
Crisis and Emergency Risk Communication model	[4, 57, 58]
Extended parallel process model	[59, 111]
Fuzzy-trace theory	[40]
Hermann's crisis model	[87]
Narrative Policy Framework	[77, 95]
Narrative Risk Communication Framework	[77]
Precaution adoption process model	[35]
Prospect theory	[102]
Protection motivation theory	[3, 14]
Rhormann's risk communication process model	[89]
Self-efficacy design framework	[14]
Social media analytics framework	[19]
Social-mediated crisis communication model	[116]
Unspecified or combined frameworks	[16, 44, 57–59, 73, 87, 101, 115]

Note: Some articles addressed or combined multiple frameworks; some frameworks were not specified explicitly.

Computational Methods and Message Construction Efficiency

No studies in this SLR addressed using computational methods to improve message construction efficiency. Any coverage of efficient computational methods analyzed and compared various machine learning models on how well risk messages were classified based on framework or hazard keywords [2, 62, 97]. Support vector machines (SVMs), Extreme Gradient Boost (XGB), Naïve Bayes, and random forest were commonly used for efficient classification of message elements (Table 3.5).

Trade-offs Between Method Efficacy and Efficiency

No articles analyzed trade-offs between efficacy and efficiency in computational message construction. Limitations with effective and efficient computational methods were rarely discussed as well. Rather, discussions mainly focused on a lack of NLP term dictionaries for low-resource languages [2] and NLP tools inconsistently analyzing sentiment in text [19].

Message Classification with Computational Methods

Classification of messages was the predominant application of computational methods in RCC. Computational methods helped identify and classify message elements that were effective in changing individuals' risk perceptions, mitigation behavior, and self-efficacy [2–4, 16, 44, 57–59, 62, 73, 97, 101]. The most prevalent computational methods used for message and element classification were NLP, content analysis, logistic regression, Naïve Bayes, SVMs, and XGB (Table 3.5). All the above computational methods contributed toward improving the efficacy and efficiency of message classification.

Operationalized theoretical frameworks were also used to classify messages on their efficacy [4, 16, 40, 57–59, 73, 101, 116]. The CERC model was the most common theoretical framework operationalized for message classification. Combined or unspecified frameworks were more common in message classification studies than in studies that used computational methods to analyze or construct RCC messages. These combined or unspecified frameworks were covered in nine articles [16, 44, 57–59, 73, 87, 101, 115] (Table 3.6).

Discussion

We report current and emerging aspects in RCC research that improve message construction. Discussions on specific methods used to improve message construction efficacy and efficiency, as well as their limitations, are also present in each subsection below. We also share insight into how cultural differences within affected populations can impact RCC message development.

Theoretical Frameworks in Risk and Crisis Communication

Effective RCC depends on the theoretical framework chosen for computational message development. The most commonly referenced frameworks used in message construction and analysis are the NPF [77, 94, 95] and PMT [3, 14, 60], as described in Sections "Protection Motivation Theory (PMT)" and "The Narrative Policy Framework (NPF)." Choosing between the two frameworks depends on how researchers want to motivate individuals to change risk mitigation behavior and risk perception.

Protection Motivation Theory (PMT)

RCC developed using PMT targets individuals' fear-appeal when faced with a hazard through threat and coping appraisal [11, 60]. Fear-appeal in PMT messages is surmounted when an individual's response efficacy and self-efficacy outweigh the costs of taking protective action against the hazard communicated. The most common hazard domain applying PMT in messaging is cybersecurity, specifically RCC on increasing and encouraging cybersecure behaviors in users [3, 14]. Visual and interactive messaging, embedded with PMT tenets, improves user self-efficacy, response efficacy, and changes in cybersecure behavior [3, 14].

The Narrative Policy Framework (NPF)

The NPF is another theoretical framework used to improve message efficacy. The NPF asserts that narrative elements such as plot, setting, moral, and characters-in-action play an important role in the policy process [94]. Specific characters that appear in narratives are heroes, villains, and victims [94]. Hero characters can improve the efficacy of messages created using the NPF [77, 95]. Furthermore, communicating risk using narratives invokes narrative transportation [28] and makes the messaging more personable and memorable for individuals [15, 77, 95, 103]. Messages created with the NPF can heighten affective response and thereby have a greater impact on intended behavior as compared to strict science messages [77, 95].

Trade-offs Between PMT and NPF Inducing individual affective response differs among messages developed using PMT and the NPF. With PMT, *negative* affective response is influenced using fear-appeal [3]. With the NPF, *positive* affective response is influenced using character selection [77]. Risk messages that induce a positive valence of affect motivate individual risk mitigation behavior better than messages that induce a negative valence of affect; however, negative affect appears to have as much of an impact on individual risk perceptions as positive affect [77]. For crisis communications, the magnitude of affective response may be more important than the valence of affect in crisis messaging because individuals need to take protective actions promptly [3, 14]. However, in risk communications, we posit that the valence of affect may be more important if the messages are deployed frequently.

RCC developed with either framework is effective at inducing affective response. Although messages developed with PMT can induce affective responses, these messages motivate risk mitigation behavior with varying degrees of success [3, 14]. Factors that impact the success of these messages are individual risk perception of a hazard and increased self-efficacy [14]. Messages developed using the NPF, on the other hand, motivate risk mitigation behavior consistently when compared to conventional RCC messaging [77, 95].

Messaging on a specific hazard, as opposed to a generalized hazard, changes how individuals perceive associated risks. Individuals focusing their attention on one hazard at a time limits cognitive overload and is correlated with changes in perception [3, 14]. The medium through which a message is communicated also impacts risk perception in individuals. Visuals or interactive elements derived from PMT tenets improve response efficacy and self-efficacy in individuals [3, 14]. Researchers have also developed visual messaging adhering to the NPF, showing such messaging to be as effective at inducing affective responses in individuals as text-based messaging [31, 93]. By visualizing the risks of a hazard, individuals understand how a given hazard affects them, thus improving risk perception.

Effective Computational Methods in Risk Communication

Applications of and discussions on computational message construction are scant in RCC research. The research focuses on classifying or analyzing RCC messages, specifically message elements like calls to action, hazard information, and visual media [2–4, 16, 19, 35, 40, 57–59, 62, 73, 87, 97, 101, 102, 111, 115, 116]. Sometimes, articles include discussions on how analyzed elements improve message construction in the future, but how to construct messages is never specified [3, 16, 19, 35, 44, 57–59, 62, 73, 111]. Failure to specify how messages are constructed represents a larger problem in message development as a whole, not just computational message construction [83].

Natural Language Processing (NLP) Little research delves into using computational methods to improve text-based messaging [83]. When message construction methods are discussed, messages are constructed either with visual or video elements and compared against textual risk messaging [14, 80, 89, 103]. From the limited body of work, computational methods used in text-based message construction are linguistic computer science tools such as Linguistic Inquiry and Word Count [44] or NLP [77, 95].

NLP proves to be effective at improving message development. NLP tools have been applied in operationalizing theoretical frameworks [77, 83, 95] and analyzing message efficacy through audience engagement [19, 58, 59]. Messages developed using NLP can automate content analysis [70, 77, 95] and help select terms associated with framework elements [77, 83, 95]. Additionally, NLP tools such as sentiment analysis can help determine how individuals respond to RCC messaging by examining the sentiment expressed in RCC messages [19, 58, 59].

Limitations with NLP Although NLP is a powerful tool for textual analysis, limitations emerge with text classification, sentiment analysis, and topic modeling techniques. All three techniques have difficulties assessing term sentiment, classification, and topic relevance of context-specific words or sentences [19, 32, 83, 99]. In contrast, manual classification and sentiment analysis can contextualize words better than equivalent NLP tools because humans can intuit situational context better than computers [32, 99]. Difficulties in contextualizing terms can be attributed to generalized word dictionaries used when working with NLP tools.

Sole reliance on generalized word dictionaries contributes to the disparity between manual text analysis and NLP tools. For example, sentiment analysis performed on RCC messages sometimes calculate a negative polarity score (i.e., a message is interpreted to have a negative sentiment) for the whole message even if terms used are "neutral" for the given context [19]. Additionally, Part of Speech (POS) tagging with generalized word dictionaries often results in poor accuracy. Typical problems include distinguishing different noun types, ignoring multiword units, and assigning wrong POS tags [99]. Nuances of human language are challenging for computational methods to fully analyze or classify terms, a problem further exacerbated when applying NLP on low-resource languages [20].

NLP lacks substantial support for low-resource languages [25]. Typical strategies take messages written in a low-resource language (e.g., Arabic) and translate them into a high-resource language (e.g., English) before applying word classification, term frequency analysis, or sentiment analysis [2, 20, 25]. Translating the original text often loses the context or meaning of the message [25]. However, this is not the only challenge present with NLP for low-resource languages. These languages can contain linguistic and semantic ambiguities, and some do not adhere to punctuation or capitalization rules present in high-resource languages [20, 25]. Messages deployed lose their effectiveness if the wrong words are chosen or if the language structure is incoherent for message recipients.

Efficient Computational Methods in Risk and Crisis Communication

RCC research largely overlooks applications of efficient computational methods in message development. Current applications of computational message construction, while improving construction and message efficacy, are time- and resource-consuming [82]. Hence, researchers focus their attention toward large language models (LLMs) for efficient message construction [43, 56, 82]. Current work on constructing messages using LLMs combines zero-shot learning and prompt engineering to develop accurate, quality, and impactful messaging [22, 51, 55, 56]. Additionally, advancements in generating non-textual, multimedia forms of communication through LLMs are also ongoing [65, 66].

Limitations with Efficiency LLMs have become popular and powerful tools for text generation and communication research [41, 56]. However, LLMs have technical and ethical limitations with respect to accountability, responsibility, safety, and honest use [56, 88, 104, 107]. LLMs perform well when fed multimodal [106] and validated information [26, 42], and LLM prompts created with explicit guidance can result in consistent, well-structured outputs [22]. Although these methods are implemented to mitigate concerns over validity, uncertainty, bias, and accountability when generating LLM outputs, challenges still present themselves. On the technical side, generating multiple messages utilizing the same prompt can result in temporal mismatches within the message content across a set of messages [56]. Issues also arise when irrelevant contextual or personal information is introduced into the prompt [55, 98]. On the ethical side, LLMs can generate messages embedded with social, scientific, and psychological biases [33, 112] or provide morally inconsistent advice [45]. In addition, responses given by popular LLM ChatBots, such as OpenAI's ChatGPT[12],

[12] https://chat.openai.com/

can contain inaccurate or overgeneralized information, stemming from limited or biased training sets [107]. However, solutions are being explored to mitigate the impacts of position, verbosity, and self-enhancement biases [117].

These limitations with LLM message generation threaten the objectivity and accuracy of RCC messaging. Therefore, it is important that researchers take full consideration of the trade-offs between timely message deployment and precise message content. With these concerns in mind, our position is that message construction should not be fully reliant on LLMs. Rather, LLM-generated messages should involve human validation to ensure message content is accurate [55] and to judge whether linguistic nuances are properly reflected [69].

Impact of Culture on Risk and Crisis Communication Messaging

Computational tools and theoretical frameworks are instrumental for improving RCC message construction efficacy and efficiency. However, variation in cultural contexts impacts the efficacy of computational tools. A significant cultural barrier is the predominance of the English language in computational linguistic tools and dictionaries [2, 20, 25]. For all other languages, the tools' reliance on English requires that text be translated. Because translation can ignore or introduce semantic ambiguities, we posit that RCC messages will be less effective if they require translation. Consequently, we expect that linguistic barriers reduce the efficacy of messages when dispersed to non-English-speaking populations. Therefore, language is an important cultural consideration.

Language is not the only cultural barrier that impacts RCC messaging. Culture is inextricably linked to geolocation. For example, Asian countries tend to be more collectivist than Western countries like the United States [111, 113]. Collectivism is an example of a cultural factor that can impact message receptivity. Message receptivity is also impacted by intercultural differences such as race, ethnicity, political affiliation and ideology, cultural norms and prevailing personal beliefs and attitudes (e.g., religious and philosophical) [14, 77, 93, 113]. Therefore, cultural influences on message receptivity cannot be tackled with computational methods alone because linguistic and other cultural considerations influence message efficacy.

Threats to Validity

Construct Validity

We identified construct validity as a potential threat. Construct validity refers to the extent to which an instrument or test reflects the construct being investigated [83]. Our investigation into computational message construction methods involved both manual and automatic filtering of literature. The use of manual filtering methods can threaten construct validity. Both the abstract reviews and quality assessments used manual filtering methods via reading key findings of a given article and recording criteria for SLR inclusion. The inclusion criteria served as a guide for assessing article relevance. Yet, differences in how reviewers assessed relevance could have introduced some inconsistencies in the inclusion of articles meeting the inclusion criteria in Table 3.3 by a slim margin.

We selected reviewers that had a solid knowledge base on computational sciences to ensure literature on computational RCC was included in the SLR. Of the six reviewers, all were familiar with RCC, but only half of them were RCC experts. Although it would have been better if all reviewers had been experts in RCC, our stringent and well-defined inclusion criteria mitigated threats to construct validity resulting from some of the reviewers' limited expertise in RCC.

In addition, we did not conduct overlapping abstract reviews, meaning distributed articles were unique for each reviewer. Distributing articles with reviewer overlap would have improved the reliability and construct validity of the study. However, the large volume of articles (Figure 3.2) was a significant undertaking for our team; hence, our decision to not distribute overlapping articles. We perceive threats resulting from this to be limited to articles that were included or excluded by a narrow margin, as mentioned earlier in this section.

Computational filtering methods may have also introduced threats to construct validity. The TF-IDF analysis potentially threatens construct validity by assuming frequently occurring terms correlate with an article's relevancy to computational risk message construction. Most frequently occurring terms across all vetted articles were not deemed research-relevant, so it is likely that research-relevant articles were falsely excluded or included.

Another potential threat to construct validity stems from the database searches, specifically with the search string construction. The strings went through a month-long refinement process to best capture relevant literature on RCC messaging. However, it is possible that the search strings did not capture all relevant literature on the topic. For example, the studies of [83] and [56] were sources not captured in our SLR despite their discussion of computational RCC message construction. The search strings also yielded some results irrelevant to RCC messaging, which we manually filtered out both in the database search and results screening steps of this SLR (Figure 3.2).

External Validity

External validity refers to whether results from this study can be generalized beyond the specific research content [12]. We searched for literature on RCC messaging across four distinct databases. For each database, we limited our search to include literature published between the years 2018 and 2023. This range was chosen to best capture any RCC research covering LLMs or NLP. However, our results from this research scope could be too specific to generalize to most RCC research available.

Conclusion

The primary application of computational methods in RCC is for message classification. Computational methods help classify effective RCC message elements, and similar methods classify RCC messages based on hazard domains efficiently. However, computational methods are seldom used to improve the efficacy and efficiency of message construction. Although some RCC research leverages computational methods to improve message construction efficacy, improving construction efficiency is a nascent area of research but one that is rapidly growing with the maturation of LLMs. Yet, by improving message construction efficacy *and* efficiency, RCC messaging would have greater potential to quickly warn and better inform affected communities impacted by hazards. Thus, we recommend that future RCC research focus on the development of computational methods for improving *efficacy and efficiency* in message construction.

Acknowledgements

The authors thank A. Redempta Manzi Muneza and Tom McElroy from the MSU Software Engineering and Cybersecurity Lab (SECL) for their assistance with supplemental abstract reviews. We also thank Seyedmojtaba Mohasel from the MSU Department of Mechanical Engineering and Garrett Perkins and Zach Wadhams from the MSU SECL for providing insightful and constructive feedback on earlier drafts of this article.

Conflict of Interest Statement

The authors declare no conflicts of interest.

INTEGRATING COMPUTATIONAL TEXT ANALYSIS IN RISK AND CRISIS COMMUNICATION DEVELOPMENT

Contribution of Authors and Co-Authors

Manuscript in following chapter

Author: Madison H. Munro

Contributions: Developed study concept and design, data collection and analysis, interpretation of results, and wrote the manuscript.

Co-Author: Manuel Ruiz-Aravena

Contributions: Developed initial program for content analysis, provided feedback on manuscript and code, and assisted with editing.

Co-Author: Elizabeth A. Shanahan

Contributions: Assisted with data analysis validation and human subjects research documentation, provided feedback on manuscript, and assisted with editing.

Co-Author: Savanna Washburn

Contributions: Assisted with data analysis validation, provided feedback on manuscript, and documented human coding steps.

Co-Author: Ann Marie Reinhold

Contributions: Provided general study direction, feedback, and substantial edits on the manuscript.

Manuscript Information

Madison H. Munro, Manuel Ruiz-Aravena, Elizabeth A. Shanahan, Savanna Washburn,

Ann Marie Reinhold

2025 IEEE International Conference on Information Reuse and Integration and Data Science (IRI), San Jose, CA, USA

Publisher: IEEE
In-press
DOI: 10.1109/IRI66576.2025.00075

Status of Manuscript:
____ Prepared for submission to a peer-reviewed journal
____ Officially submitted to a peer-reviewed journal
____ Accepted by a peer-reviewed journal
__X__ Published in a peer-reviewed journal

Abstract

Qualitative textual analysis is advancing with the integration of Natural Language Processing (NLP) and Large Language Models (LLMs). Although many multidisciplinary researchers adopt these analysis tools, Risk and Crisis Communication (RCC) researchers have not taken full advantage of such tools in content analysis tasks. We investigate computational tools' ability to replicate human analysis in RCC research. We integrate NLP and ChatGPT into content analysis steps performed on insider threat source text. The first content analysis step tasks ChatGPT with coding insider threat text as character types defined by the Narrative Policy Framework (NPF), and the second step leverages select NLP tools to derive meaning in the coded sentences. Content coding with ChatGPT varies in performance based on the amount of detail present in prompts. Select NLP tools have varying degrees of success when processing and analyzing coded text. Both ChatGPT and NLP require human intervention when performing content analysis, thus a mixed method approach is necessary for future RCC research.

Introduction

Advancements in computer science provide empirical techniques to process and analyze large sets of text. These techniques include Natural Language Processing (NLP), a set of machine learning tools that help with text classification and analysis [83], and Deep Learning models like Neural Networks and Large Language Models (LLMs) [46]. Both NLP and Deep Learning are widely adopted by researchers across multiple research domains– from STEM and medicine to the humanities and social sciences [105, 110].

Computational textual analysis enables researchers across various social science fields to empirically analyze qualitative text [71, 83]. One research domain where computational textual analysis is instrumental is Risk and Crisis Communication (RCC). A handful of RCC researchers have integrated NLP and LLMs to swiftly develop effective messaging across various hazard domains [43, 51, 83]. However, the use and integration of NLP and LLMs remain exploratory in RCC research.

Development of RCC messaging typically relies on manual content analysis [67, 83]. This content analysis is guided by a theoretical framework for identifying key terms and phrases, message structure, and beliefs expressed [67]. Solutions for guiding content analysis with computational tools have been proposed and tested [67, 77, 83, 95]. Moreover, messaging strategies are needed that bridge across hazard types and scientific disciplines and domains [81]. One solution proposed is the Domain Agnostic Risk Communication (DARC) Framework [82].

The DARC Framework integrates computational textual analysis in RCC message development to systematically improve content analysis and message construction efficacy [82]. Furthermore, the DARC framework reduces RCC development time through the integration of LLMs, thus enabling timely message delivery to impacted populations. We apply this framework to develop RCC for organizational insider threats. We focus specifically on steps using NLP and LLMs to perform content analysis, using the Narrative Policy Framework (NPF) [94] as the guiding framework.

The following research questions (RQs) drove our insider threat RCC content analysis:

RQ1. Can LLMs perform content coding[1] as good or better than human coders using the NPF as a guiding framework?

RQ2. Can select NLP techniques effectively identify and analyze NPF language to include in future RCC messaging?

Methods

Gather Insider Threat Source Text

Insider threat source text came from two places: structured interviews and authoritative sources. For interviews, we recruited eleven participants from five organizations via snowball sampling (IRB Protocol #2024-1486-EXEMPT). Each participant varied in insider threat expertise, ranging from cybersecurity experts to individuals with minimal computational skills. We interviewed participants on their experiences dealing with organizational insider threats; three interview questions targeted Problem Definitions, three targeted Risk Perceptions, and three targeted Solutions. Responses to each question were recorded and then transcribed.

Authoritative insider threat sources were pulled from three sources: the USA Cybersecurity & Infrastructure Security Agency's (CISA) 2020 Insider Threat Mitigation Guide[2], Gurucul[3,4], and Montana State University's Insider Threat Mitigation Guide. Each

[1] The categorization of language in qualitative source texts (*sensu* [83])

[2] https://www.cisa.gov/resources-tools/resources/insider-threat-mitigation-guide

[3] https://gurucul.com/blog/what-is-an-insider-threat/

[4] https://gurucul.com/blog/best-insider-threat-tools-and-strategies-for-mitigating-risks/

source was grouped into one of three hazard information sections—Problem Definition, Risk Perception, and Solutions—then underwent text preprocessing.

Data Preprocessing

Both sets of insider threat source text underwent manual and computational preprocessing. For manual preprocessing, we redacted sensitive interview information, extracted interview responses, and removed figures and tables within authoritative source text. For computational preprocessing, we converted the text to lowercase and removed numbers, URLs, and punctuation from the text (except for ending punctuation). All computational preprocessing was done in *R* [76] version 4.4.3 using the *tidytext* [100] package.

The preprocessed text was split into two copies before content coding steps. The first copy contained three subsets corresponding to hazard information sections defined in section 4, and the second copy contained the original text separated into individual sentences. Both copies provided data for subsequent content coding steps with OpenAI's ChatGPT model[5].

Content Coding with ChatGPT

Content coding on insider threat text required two sets of instructions[6]. Both instruction sets contextualized the insider threat text, provided NPF definitions for character types, and provided explicit instructions for response generation and formatting. The instruction sets guided two separate executions of ChatGPT (model version: *gpt-4o-mini*; model temperature: *0.3*). These model instances ran within OpenAI's Python API[7] to code for characters and character language on insider threat text.

The first ChatGPT model instance identified characters within the section-level text that matched the NPF definitions for Hero, Victim, or Villain (*sensu* [94]). For each section,

[5] https://openai.com/chatgpt/overview/

[6] Appendix A: https://doi.org/10.5281/zenodo.15027281

[7] https://openai.com/api/

we ran the model three times to code for the three character types. All characters identified were fed into the next model instance as a list of potential insider threat characters.

The second ChatGPT model instance classified sentence-level text as Hero, Victim, or Villain language. Language groupings were based on which NPF-defined characters frequently appeared in a sentence and how these characters were framed in the sentence. Characters came from the list compiled in the first ChatGPT model instance.

The second model instance used three different prompts for the classification task. The first prompt classified sentences as (1) one or more NPF character types, (2) "NONE" if the sentence contained no characters, or (3) "UNCLEAR" if the character types were unclear. The option to label sentences as "NONE" was removed in the second prompt, and the third prompt removed the option to label sentences as "NONE" or "UNCLEAR". All prompt information was otherwise identical. These three different versions of the classification prompts were used to assess differences in inter-coder reliability between ChatGPT and humans.

The prompt engineering and human coding were aligned with best practices in social science research. Specifically, human coders and ChatGPT were allowed to assign multiple character labels to each sentence.

Content Coding Inter-coder Reliability

Content coding with ChatGPT underwent two rounds of inter-coder reliability to assess character type label agreement. The first round assessed the inter-coder reliability between a subset of ChatGPT-coded data and the same data subset coded by two trained human content coders. The data subset consisted of a 20% stratified random sample (SRS) of the section level-data and a 10% SRS of the sentence-level data. To code each SRS, the human coders followed the same instructions developed for ChatGPT content coding, and the coders adhered to the first prompt for sentence classification[8]. Inter-coder reliability between ChatGPT and the human coders was measured via Cohen's Kappa [64].

The second round assessed the inter-coder reliability between two copies of both the full section- and sentence-level text. For the sentence-level text, we used the classification prompt that exhibited the best inter-coder reliability between ChatGPT and the human coders. To assess inter-coder reliability between the two copies of sentence-level text, we ran the content coding pipeline described in Section 4.2.3 twice; inter-coder reliability between the two texts was measured via Cohen's Kappa for this round as well. After measuring Cohen's Kappa, we merged the two texts based on their assigned character label. Any disagreements in character label were resolved by assigning the sentence a multi-character label.

Natural Language Processing Tasks

The final coded sentence-level text was copied and grouped into two corpora sets for NLP analysis. One corpora set grouped sentences based on assigned character type, and one corpora set grouped sentences based on hazard information section. Both corpora sets provided text input for Frequency Analysis and Sentiment Analysis, and the original copy of the sentence-level text provided a dataset for Word Classification.

Word Classification We performed Word Classification in *Python* version 3.12.5.[9] to predict character labels for insider threat sentences. Two classification models were fit to an 80/20 training and testing dataset of the coded sentence data. The models chosen were a Random Forest model and a Support Vector Machine (SVM). We trained both models with the training dataset encoded by the BERT [18] model "all-mpnet-base-v2." After fitting both models and encoding the training dataset, we assessed the accuracy, precision, recall, and F1 scores of character classification on the testing dataset.

[8] Appendix B: https://doi.org/10.5281/zenodo.15027281

[9] https://www.python.org/

Frequency Analysis We assessed the frequency of insider threat characters within hazard sections by calculating their Term Frequency-Inverse Document Frequency (TF-IDF)[10]. Before calculating TF-IDF, we removed stopwords from sentences using both a custom stopword library and the *smart*[11] *R* stopword library. We then calculated the TF-IDF for uni-, bi-, and trigrams within each sentence. For each ngram investigated, we also calculated the TF-IDF for their stemmed equivalents. The stemmed and nonstemmed ngrams provided a wider range of words from which meaning was derived.

Sentiment Analysis We measured the sentiment of insider threat sentences in each corpora set. For each sentence in each corpora set, we calculated the overall sentence sentiment by summing the polarity scores calculated for each term in the sentence. Term polarity was determined using polarity scores provided by the *afinn*[12] library in *R*. After calculating the sentence sentiment for each corpora set, we performed One-Way Analysis of Variance (ANOVA) on the data, assessing any differences in mean sentiment across character and section corpora. We also performed Two-Way ANOVA to assess differences in mean sentiment across character types within sections.

Results

Character Coding Inter-coder Reliability

The inter-coder reliability between ChatGPT and human coders varied depending on which sentence classification prompt was used. As the level of detail in each prompt decreased, label agreement between ChatGPT and the human coders increased (Table 4.1). In contrast, the inter-coder reliability between the two ChatGPT coded data indicated very high agreement in character label assignment. Assessing label agreement yielded a Cohen's

[10] https://cran.r-project.org/web/packages/tidytext/vignettes/tf_idf.html
[11] https://cran.r-project.org/web/packages/stopwords/readme/README.html
[12] https://afit-r.github.io/sentiment_analysis

Table 4.1: Cohen's Kappa Calculated for Character Label Agreement Between ChatGPT and Human Coders.

Sentence Classification Prompt Used by ChatGPT	**Cohen's Kappa**
Classify sentence as one or more NPF character type, "NONE", or "UNCLEAR"	0.309
Classify sentence as one or more NPF character type or "UNCLEAR"	0.393
Classify sentence as one or more NPF character type	0.474

Table 4.2: Performance Metrics of the Random Forest Model and SVM on the 80/20 Dataset Split.

Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**
Random Forest	0.393	0.388	0.393	0.356
SVM	0.398	0.392	0.398	0.377

Kappa of 0.907.

Word Classification

Both Word Classification models classified insider threat sentences as Hero, Victim, or Villain language with low fidelity (Table 4.2). The Random Forest model averaged a 0.321 model performance accuracy after 10-fold cross-validation. These cross-validation scores ranged from 0.158 to 0.629 (Fig. 4.1).

The SVM performed better than the Random Forest model on average (Table 4.2). The SVM averaged a 0.438 model performance accuracy after 10-fold cross-validation. These cross-validation scores varied less than scores for the Random Forest model, ranging from 0.333 to 0.703 (Fig. 4.1).

Frequency Analysis

Frequency Analysis on the section corpora unveiled potential key insider threat characters and terms. Nonstemmed bigrams revealed clearer characters and character

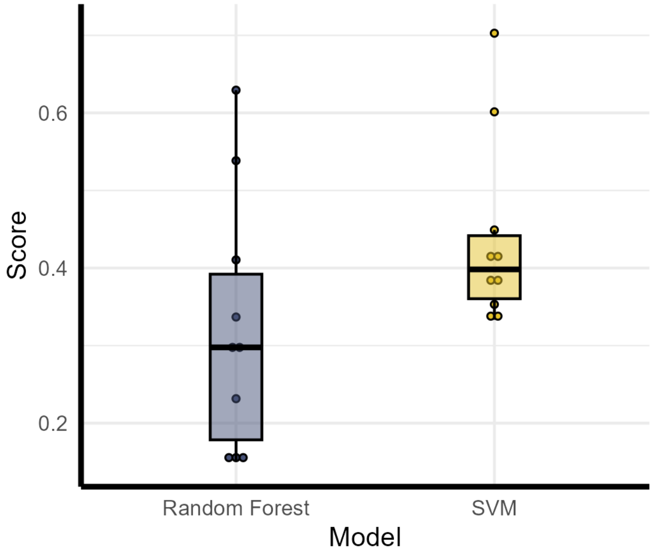

Figure 4.1: Comparison of 10-fold cross-validation accuracy scores between the Random Forest and SVM models.

language compared to other stemmed and nonstemmed ngrams investigated[13].

Character types impacted which insider threat bigrams were important within hazard information sections (Table 4.3). Many of the bigrams in the section corpora corresponded to explicit Hero, Victim, and Villain characters. Some of these bigrams show up as more than one character type (e.g., the bigram "compromised data" in the Risk Perception corpus). Other bigrams corresponded to character language as opposed to actual characters (e.g., the bigram "feel educated").

Sentiment Analysis

Sentiment Analysis provided insight into how character types impact insider threat sentence structure (Fig. 4.2). Within the character corpora, mean sentiment differed significantly across each corpus (One-Way ANOVA, F-score: *6.592*, p-value: *0.001*, df: *2*). Within the section corpora, there were no detectable differences in mean sentiment across

[13] Appendix C: https://doi.org/10.5281/zenodo.15027281

Table 4.3: Top Three Character Type Bigrams across the Section Corpora based on TF-IDF.

Character Type	**Problem Definition**		**Risk Perception**		**Solutions**	
	Bigram	***TF-IDF***	***Bigram***	***TF-IDF***	***Bigram***	***TF-IDF***
Hero	“indicator person”	0.251	“compromising data”	0.365	“quality code”	0.106
	“phishing emails”	0.147	“vigilant groups”	0.293	“resourcing people”	0.071
	“flagged organization”	0.139	“cyber threats”	0.221	“access information”	0.070
Victim	“negligent data”	0.120	“compromising data”	0.164	“job position”	0.180
	“organization logs”	0.120	“feel educated”	0.147	“resourcing people”	0.115
	“information leaving”	0.115	“cyber attacks”	0.139	“research lab”	0.092
Villain	“indicator person”	0.139	“compromising data”	0.453	“backing things”	0.264
	“indicators compromise”	0.122	“delete stuff”	0.377	“taking extra”	0.133
	“adept technology”	0.120	“high risk”	0.165	“drives backing”	0.132

sections (One-Way ANOVA, F-score: *0.016*, p-value: *0.984*, df: *2*).

Mean sentiment did not differ significantly when only considering sections (Two-Way ANOVA, F-score: *0.038*, p-value: *0.963*, df: *2*). However, mean sentiment differed significantly across each section when accounting for character types (Two-Way ANOVA, F-score: *6.694*, p-value: < *0.001*, df: *2*).

Discussion & Future Work

Content Coding with Large Language Models

With respect to RQ #1, leveraging LLMs like ChatGPT can streamline qualitative content coding steps in RCC development. Our content coding results show that ChatGPT has near complete agreement in character label assignment (above 80%) between separate instances of itself. Achieving high label agreement indicates that at least 80% of the coded data is well-represented, thus conclusions drawn from the data are sound [64]. Coding data with ChatGPT suitably replicates content coding while ensuring coded data is well-represented.

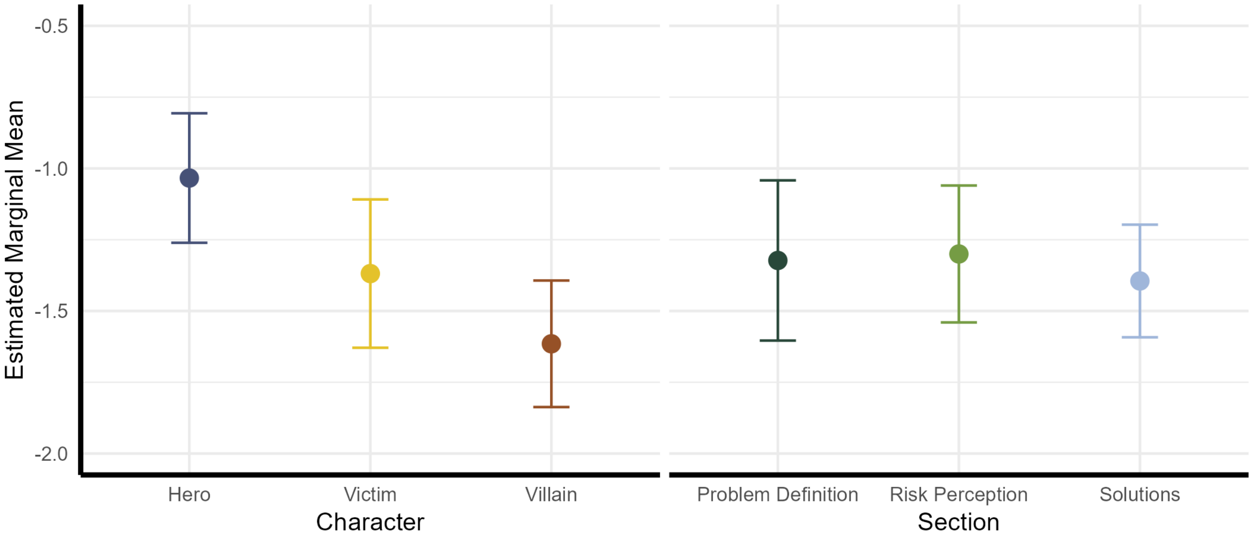

Figure 4.2: Mean Sentence Sentiment Across Character Types (left) and Across Hazard Information Sections (right).

Replicating the time-consuming and laborous task of content coding is achieved through robust prompt engineering. Creating robust prompts combines zero-shot or few-shot learning with well-structured response generation instructions [38, 61]. Our prompts for NPF content coding combine few-shot learning with structured prompt components such as study context, explicit instructions, and output format. Robust prompt engineering can lead to improved performance in qualitative text analysis [78]. However, qualitative text analysis done by a well-prompted LLM requires human involvement.

Content coding using an LLM needs human intervention to assist in validation and quality assurance. While LLMs can be refined to return high quality results and performance, prompt engineers risk over-fitting the model [13]. Our inter-coder reliability results for ChatGPT and human coders indicate that including more detail in coding instructions returns lower quality responses (Table 4.1). Additionally, LLMs cannot use deductive reasoning like humans, leading to less nuanced responses compared to humans [6]. Responses generated by LLMs can also be heavily influenced by biased or inaccurate information [67].

To ensure accurate, nuanced, and unbiased information is captured in LLM responses, future integration of LLMs in RCC research should include humans.

Content Analysis with Natural Language Processing

With respect to RQ #2, NLP techniques leverage quantitative analysis to derive meaning in qualitative source text. Our NLP analysis indicates that Sentiment Analysis and Frequency Analysis reveal NPF characters and character language that could have been missed by manual content analysis (Table 4.3, Fig. 4.2). Both NPF characters and character language determine the narrative structure and hazard content needed for effective RCC messaging [83, 94]. However, the approach we used for Word Classification falters when determining narrative structure in hazard source text (Table 4.2, Fig. 4.1). Specifically, allowing for multiple character labels for each sentence (Section 4.2.3) negatively affects the Word Classification models' performance. We conducted additional preliminary analysis in an effort to improve performance, and found that allowing only one character label per sentence increases the performance of Word Classification. Nevertheless, determining message structure, content, and language requires researchers to provide extra contextual information and to remain involved in content analysis [71, 83].

Natural Language Processing cannot fully replace human involvement in content analysis steps. Content analysis with NLP is constrained by the robustness of NLP tools and techniques [67, 71]. All NLP techniques investigated in this study vary in efficacy depending how text is vectorized for classification tasks [92], how researchers tokenize, stem, and preprocess text [67] and what specific models are used to classify text [71] or determine text sentiment [63]. Furthermore, leveraging high-performing NLP techniques may only *augment* qualitative content analysis rather than automating it [83]. Thus, the union of qualitative human and quantitative NLP content analysis would lead to more effective RCC message development.

Future Directions for Computational Content Analysis

Integrating *quantitative* tools to analyze *qualitative* data is enhanced by the involvement of RCC researchers. To that end, RCC researchers benefit from adopting a mixed methods approach when developing RCC messaging [82]. Adopting a mixed methods approach like the DARC Framework enables RCC researchers to develop effective and efficient messaging [82].

Developing effective RCC messages efficiently is a core function of the DARC Framework [82]. Our current work covers applying the DARC Framework to content analysis steps—a critical pre-requistite for developing future RCCs—while also integrating ChatGPT in these steps. We find that ChatGPT is limited when integrated into content coding. For future content coding tasks, it would be beneficial to (1) investigate GPT models optimized to return more nuanced and unbiased responses, (2) simplify character label classification, or (3) investigate other LLMs such as Meta's Llama model[14], which allows researchers to specify the model's training data to generate accurate responses. Using a different LLM may perform content coding as well—or better—than humans, thus improving other content analysis steps in RCC message development.

Threats to Validity

Construct Validity

Construct validity refers to the extent our study assesses the investigated construct [83]. We identified some potential threats to construct validity with our preprocessing steps, namely determining hazard information sections and redacting text. Determining if text fell under the sections Problem Definition, Risk Perception, or Solutions was guided by NPF experts. Regarding the redacted text in the interview transcripts, redacting certain text was

[14] https://www.llama.com/docs/overview/

necessary given the sensitivity of the information relayed.

We identified a few threats to construct validity with our implementation of ChatGPT for content coding tasks. First, the human coders did not code for the whole data. Performing human coding on the full source text would have been time-consuming and expensive for our team, thus the coders coded on a randomly selected subset of data. The second threat identified is ChatGPT's high variability in responses. We implemented several measures to limit as much response variability as possible. These measures include providing stringent, well-defined instructions in response generation prompts [61], lowering model temperature to reduce the likelihood of model hallucination [8], and using the most advanced model version available [49].

Internal Validity

Internal validity refers to the extent to which our cause-and-effect conclusions are sound [83]. We found no major threats to internal validity with our ANOVA on the data. Before performing ANOVA on the data, we assessed threats to statistical assumptions of equal variance of residuals, normality, significant outliers, and independence of observations. We found no significant violations of the above assumptions in our evaluation. Additionally, there was no statistically significant difference between fitting an interactive model for the Two-Way ANOVA versus an additive model (ANOVA model comparison, F-score: *1.198*, p-value: *0.310*, df: *4*). Thus, we fit an additive model for the Two-Way ANOVA.

External Validity

External Validity refers to whether our study results can be generalized beyond the current research [67]. The use and integration of LLMs and NLP in RCC research is well documented [67, 71, 83] and can be applied to future RCC research. However, our current research uses the NPF as the guiding framework for content analysis and insider threats as the domain of interest, which could be too specific to generalize to most RCC research on its own. Yet, our application of the DARC Framework ensures that content analysis and subsequent message construction can be generalized to any hazard or theoretical framework of choice [82].

Conclusion

Risk and Crisis Communication research benefits from a mixed methods approach for message development. Key components for effective and efficient RCC message development are the use and integration of NLP and LLMs in content analysis. However, each content analysis step continues to benefit from human validation and quality assurance. Human intervention in content analysis will likely be reduced as NLP tools and LLMs advance and refine. Continued and cogent integration of NLP and LLMs into RCC development will improve the efficacy and efficiency of future RCC messaging.

Acknowledgments

This human subjects research was approved by the Montana State University IRB as protocol #2024-1486-EXEMPT. M. Ruiz-Aravena was supported by the U.S. Forest Service International Programs (Agreement 22-DG-11132762-347) and the work is a contribution of the Forest and Wildlife Research Center, Mississippi State University, supported by McIntire-Stennis funds. We thank Dr. Clemente Izurieta and Brittany Boles from the Montana State University (MSU) Software Engineering & Cybersecurity Lab (SECL) for providing expertise in NLP analysis steps and providing critical feedback on this manuscript's writing and structure, respectively.

PERSUASION WITH MORE PRECISION: LEVERAGING LARGE LANGUAGE MODELS TO IMPROVE INSIDER THREAT RISK COMMUNICATION

Contribution of Authors and Co-Authors

Manuscript in following chapter

Author: Madison H. Munro

Contributions: Developed study concept and design, LLM prompt engineering, interpretation of results, facilitated validation steps, and wrote the manuscript.

Co-Author: Quinn J. Lue

Contributions: Assisted with manual validation, provided study feedback, and assisted with edits to the manuscript.

Co-Author: Ann Marie Reinhold

Contributions: Provided general study direction, feedback, and substantial edits on the manuscript.

Manuscript Information

Madison H. Munro, Quinn J. Lue, and Ann Marie Reinhold

Publisher: Journal of Military Cyber Affairs

Status of Manuscript:
____ Prepared for submission to a peer-reviewed journal
_X_ Officially submitted to a peer-reviewed journal
____ Accepted by a peer-reviewed journal
____ Published in a peer-reviewed journal

Abstract

Insider threats are an ever-increasing cybersecurity hazard. Depending on the scale and severity of the threat, organizations can face extreme financial and infrastructural costs, leaving them vulnerable to future cyberattacks. To be more proactive toward mitigating future threats, organizations can develop and deploy risk communication messaging. Efficient development of effective risk communication is crucial for insider threat mitigation. However, most risk communication research follows time-consuming black box approaches to message development. We evaluate how efficient Meta's *Llama* model can generate insider threats risk messages while promoting instrument fidelity. We also evaluate the generated content's ability to adhere to the Narrative Policy Framework (NPF) when framing insider threat risk messages. We find that *Llama* generates content efficiently in a practical sense and can produce quality, accurate, and precise insider threat message content adhering to NPF structure. The performance of *Llama* can be enhanced for risk communication development tasks with robust model fine-tuning and training data selection. Organizations can become more resilient towards insider threats with the help of proactive mitigation techniques such as risk communication

Introduction

Insider threats are an ever-increasing cybersecurity hazard impacting organizations across many domains, from nonprofits to the military [29]. Many organizations face extreme financial and infrastructural damage due to insider threats[1,2]. Damage caused by insider threats can be intensified by time-, money-, and resource-consuming threat detection and prevention tools, often implemented reactively during or after an attack[1,2].

Shifting from a reactive to a proactive stance is critical, particularly for military insider threats. Key to insider threat prevention is proactive and scalable hazard mitigation rooted in human-centered approaches[1]. One proactive and scalable approach is developing risk communication about mitigating and preventing future insider threats [68, 85]. The target population is individuals who encounter insider threats in their field of work.

Communicating hazard risk effectively and efficiently is critical for insider threat mitigation. Effective risk communication motivates target populations to adopt protective actions against a hazard [15, 67, 82, 83]. In tandem with effective risk communication is efficient message delivery. With swift message delivery and deployment, affected populations are allowed more time to prepare for and respond to a hazard [54, 67]. Given the often time-sensitive nature of insider threat response, *efficient* risk message development is a necessity.

Efficient development of effective insider threat risk messaging requires the collaboration of advanced computational analysis and guided social science frameworks [82]. Many risk communication researchers advocate for this collaboration when developing effective messages [32, 68, 71, 72, 77, 83, 95]. However, very little risk communication research investigates improving message development efficiency [67]. Additionally, many risk communication researchers present black box approaches for message development, which

[1] https://ponemon.dtexsystems.com/

[2] https://www.syteca.com/en/blog/insider-threat-statistics-facts-and-figures

can compromise the study's validity [83]. Risk communication research can benefit from a systematic approach for efficient and effective message development. One proposed approach is the Domain Agnostic Risk Communication (DARC) Framework [82].

The DARC Framework guides efficient development of effective risk messaging without compromising the validity and precision of resulting messages [82]. However, there is scant application of the DARC Framework in practice. Current applications of this framework focus on operationalizing content analysis on hazard source text with computational text analysis tools [27, 68]. In particular, the analysis presented in [68] provides the needed groundwork for efficiently developing effective messages for insider threats.

The study of [68] performs content analysis on source texts about organizational insider threats using the Narrative Policy Framework (NPF) [94] as a guide. The study's results provide message content and structure to be included in subsequent insider threat risk communication for improved message efficacy. However, the study ends just before constructing insider threat messages. We continue the work presented in [68] by following steps of the DARC Framework that focus on message development and validation. For these steps, we leverage Meta's *Llama* 3.2 model[3] to efficiently generate insider threat risk messages that incorporate message content and structure identified in [68].

The following research questions (RQs) drove our insider threat message development and validation steps:

RQ1. How efficient can an LLM generate NPF-guided risk messaging on insider threats while preserving instrument fidelity?

RQ2. To what extent can LLMs generate accurate risk communication content about insider threats?

[3] https://www.llama.com/docs/model-cards-and-prompt-formats/llama3_2/

RQ3. To what extent does LLM-generated content adhere to the NPF as framing for risk communication on insider threats?

Methods

Defining Insider Threat Risk Messaging

Our insider threat risk messages adhered to the message structure outlined by the studies of [83] and [95] and validated in [77]. These messages comprised of four segments: Problem Definition, Problem Framing, Scientific Information, and Characters in Action (Table 5.1 provides definitions and examples). With Problem Definition and Scientific Information, these segments remained static and were provided by domain experts on the investigated hazard. With Problem Framing and Characters in Action, the framing of these segments changed based on which NPF-defined characters and character language (*sensu* [94]) were used; these segments determined the character type of the overall message.

For this study, the Problem Definition segment contained the insider threat definition provided by the USA Cybersecurity & Infrastructure Security Agency (CISA)[4]. For Scientific Information, we wrote the segment using certainty language [95] and utilized text from Gurucul's 2024 Insider Threat Report[5]. We developed the remaining message segments, Problem Framing and Characters in Action, by prompt engineering responses to be generated by a specialized local instance of *Llama* 3.2-3B.

Insider Threat Risk Message Construction Steps

Prompt Engineering Character-Based Risk Message Segments We developed separate prompts for the Problem Framing and Characters in Action segments. Both prompts contained some identical information such as the study context and response formatting,

[4] https://www.cisa.gov/topics/physical-security/insider-threat-mitigation/defining-insider-threats
[5] https://go1.gurucul.com/2024-insider-threat-report

Table 5.1: Insider Threat Risk Message Segments, Definitions, and Examples

Segment	**Definition**	**Example(s)**
Problem Definition	Provides a definition on the hazard anchored in material from hazard domain experts and entities [95]	An insider threat is the potential for an insider to use their authorized access or understanding of an organization to harm that organization. This harm can include malicious, complacent, or unintentional acts that negatively affect the integrity, confidentiality, and availability of the organization, its data, personnel, or facilities.
Problem Framing	Introduces characters (Hero or Victim) and identifies problems caused by the hazard [95]	**Hero:** As someone who is committed to protecting sensitive information, you need to be aware that insider threats can occur when individuals with authorized access intentionally or unintentionally compromise security. **Victim:** Insider threats can cause harm to individuals within an organization, including employees, family members, and colleagues, by compromising sensitive information, disrupting operations, and putting critical infrastructure at risk.
Scientific Information	Discusses the likelihood of when hazard will strike at some point in the future [95]	Insider threats will occur at some point in the future. Such attacks have become more frequent across many organizations of all types, increasing from 60% to 83% occurrence over the last year.
Characters in Action	Based on the character type used (Hero or Victim), the characters introduced take action against the hazard, either by providing strategies to counter the hazard or by dealing with the consequences of no preparation [95]	**Hero:** You can take action to prepare for and mitigate insider threats by working with your IT department and Cybersecurity experts to develop risk mitigation strategies, thereby preventing further damage caused by insider threats. By staying informed and vigilant about Cybersecurity best practices, you can significantly contribute to protecting yourself, your organization, and your colleagues from the negative impacts of insider threats. **Victim:** You may face harm to your family, colleagues, and the company if you fail to prepare for and mitigate insider threats, resulting in loss of sensitive information and damage to reputation. Without proper preparation, you and your organization will struggle with difficult and stressful consequences.

Note: The Problem Definition and Scientific Information segments were written by insider threat domain experts and the Problem Framing and Characters in Action segments were generated by a custom *Llama* model.

but they differed on the segment and character type. Character type definitions came from the study of [95].

We specified the character type structure for each message prompt by appending character type-specific instructions. If the character type was “Hero”, then we appended Hero-structured instructions to the prompt providing study context and response formatting; the same logic was applied to the Victim character type. Character-structured instructions included (1) information on how to generate character-specific versions of the Problem Framing or Characters in Action segment (Table 5.1), (2) a list of characters of the specified type to include—which the model would need to include at least one of in the generated response[6]—and (3) the definition of the character type for extra context (*sensu* [94]); we also embedded some examples of character type-specific segments to provide more context. Both prompts were provided to a local instance of *Llama* 3.2-3B specialized for generating insider threat message segments.

Customizing Llama 3.2-3B We pulled a local instance of *Llama* 3.2-3B from *Ollama*[7] onto a Lambda Vector desktop system (Specs: Ubuntu 22.04 LTS, AMD Ryzen Threadripper 3990X, 2x NVIDIA Quadro RTX 6000/8000, 256GB RAM). To customize the model, we created a Modelfile[8] that specified model parameters (e.g., temperature, context window size) and the system role. We specified the model with the following parameters:

- Temperature: 0.3
- Context Window Size: 4096
- Top P: 0.6
- Top K: 40

[6] The character lists came from [68]

[7] https://ollama.com/

[8] https://github.com/ollama/ollama/blob/main/docs/modelfile.md

- System Role: "Present insider threat risk communication message segment in the second person. Use lay, impersonal language"

All other model parameters adhered to the default values for *Llama* 3.2-3B. After specifying the model parameters, we created a new instance of our base *Llama* model using the Modelfile. This custom model, along with the engineered prompts, were ran in Python[9] (version 3.10.12) using the dedicated *ollama* Python package[10].

Constructing Insider Threat Risk Messages We created a pipeline within Python to construct insider threat risk communication messages. Each message was constructed by concatenating the four message segments together, starting with Problem Definition followed by Problem Framing, Scientific Information, and Characters in Action. For the Problem Framing and Characters in Action segments, we generated two versions of each section, one version structured with Hero characters and language and one version structured with Victim characters and language.

We constructed four different types of messages—Hero, Victim, Victim-to-Hero, and Hero-to-Victim—using different combinations of character type-specific segments. The Hero message contained Hero-structured Problem Framing and Characters in Action segments while the Victim message adhered to Victim structure for the same segments. For the Victim-to-Hero message, the Problem Framing segment had Victim structure, and the Characters in Action segment had Hero structure[11]. The Hero-to-Victim message had its segments structured the opposite; hero language in Problem Framing and victim language in Characters in Action. The final messages comprised of 4-8 sentence-long messages on insider threats, targeting malicious insider threats, inadvertent insider threats, or both (Table 5.1).

[9] https://www.python.org/

[10] https://github.com/ollama/ollama-python

[11] This message structure is different than the Victim-to-Hero structure specified in the studies of [83] and [95] wherein the two character type-specific segments contain both Victim and Hero language

Insider Threat Risk Message Analysis

Message Construction Runtime We measured the runtime for constructing all four message types via time trials. Each trial would run the entire message construction pipeline a set number of times, tracking the pipeline completion time. We measured the runtime of constructing 1, 5, 10, 25, 50, and 100 messages and plotted differences in runtime across message types.

Manual Validation of Constructed Messages We assessed the validity of 500 messages for each character type combination (2,000 total messages) using human validation. Two human validators assessed insider threat messages by following a set of criteria for valid messages[12]. This criteria was split into two types of criterion: risk communication-specific and general LLM response validation. Risk communication-specific criteria included adherence to the correct hazard topic of insider threats, correct use of characters and character language, and adherence to the character type specified in the message construction pipeline. General LLM response criteria included checking for response generation errors and adherence to the given prompt.

The human validators documented whether messages met the specified criteria within spreadsheets distributed by the lead author. The spreadsheets contained the messages to evaluate and their character type structure, columns aligning with the specified criteria, and a column to indicate if a given message was valid. If all criteria columns were filled out as "YES" for a given message, then the message was classified as "VALID". If any of the criteria columns contained a "NO", then the message was classified as "NOT VALID". After the message evaluation was complete, we performed computational analysis on the validation results.

[12] Appendix A: https://doi.org/10.5281/zenodo.17194612

Computational Analysis on Constructed Messages We analyzed both valid and invalid messages in Python to identify key trends with message content and structure. First, we plotted incidence of violated criteria, partitioned by character type. Next, we identified overlap of Hero and Victim characters within all messages both valid and invalid. We utilized the Hero and Victim character lists from [68] to find matches in the generated messages, then we visualized character overlap for both valid and invalid messages. Lastly, we calculated the sentiment of valid messages and their individual segments using the BERT [18] model *bert-base-multilingual-uncased-sentiment*[13]. The BERT model binned segments into one of five categories: "Very Negative", "Negative", "Neutral", "Positive", and "Very Positive". These categories were determined by the bidirectional encoding of the segment. The segments not generated by *Llama* had their sentiment calculated one time since they remained unchanged in each message.

Results

Message Construction Runtime

The message construction pipeline executed quickly across all message types. As the number of messages to construct increased, the time needed to complete the full pipeline increased linearly (Figure 5.1). Completion times varied slightly across message types. For the fastest performing message type, Hero, its runtime at 100 messages peaked at around 100 seconds (Figure 5.1). For the slowest performing message type, Victim, its runtime at 100 messages was fast in a practical sense, peaking at 120 seconds (2 minutes).

Manual Message Validation

The message construction pipeline primarily constructed valid messages. Of the 2,000 messages constructed, 70.2% of them fully met the validation criteria (Table 5.2). For both

[13] https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment

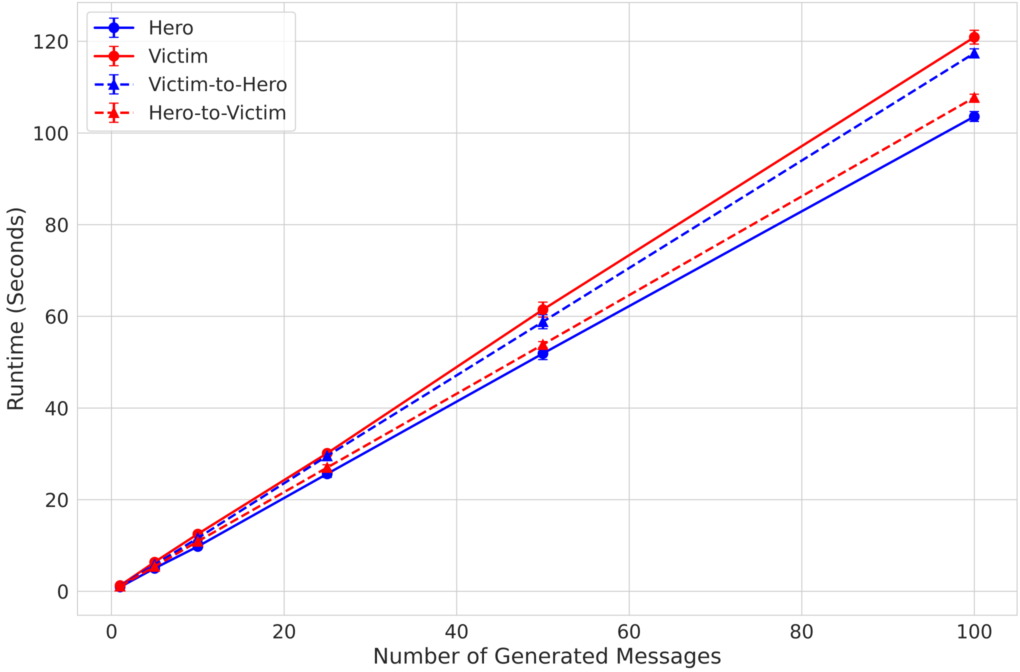

Figure 5.1: Message Construction Time Trials Across Message Type

Table 5.2: Metrics for Generated Insider Threat Risk Messages

	Hero		**Victim**		**Victim-to-Hero**		**Hero-to-Victim**		**Total**	
	Raw Count	***Percentage***	***Raw Count***	***Percentage***	***Raw Count***	***Percentage***	***Raw Count***	***Percentage***	***Raw Count***	***Percentage***
Valid Messages	365	73.0%	188	37.6%	458	91.6%	393	78.6%	1404	70.2%
Invalid Messages	135	27.0%	312	62.4%	42	8.4%	107	23.4%	596	29.8%
Raw Total	500	100%	500	100%	500	100%	500	100%	2000	100%

valid and invalid messages, the number of valid/invalid messages varied across the different message types. Victim-to-Hero messages had the highest rate of valid messages at 91.6% while Victim messages had the highest rate of invalid messages at 62.4% (Table 5.2).

The rates for violated criteria varied across message types. For instance, Victim messages violated the "Second Person" criterion more often than the other message types, yet these messages rarely violated the "Error Free" criterion unlike Hero and Hero-to-Victim messages (Figure 5.2). Victim-to-Hero messages violated the least amount of criteria.

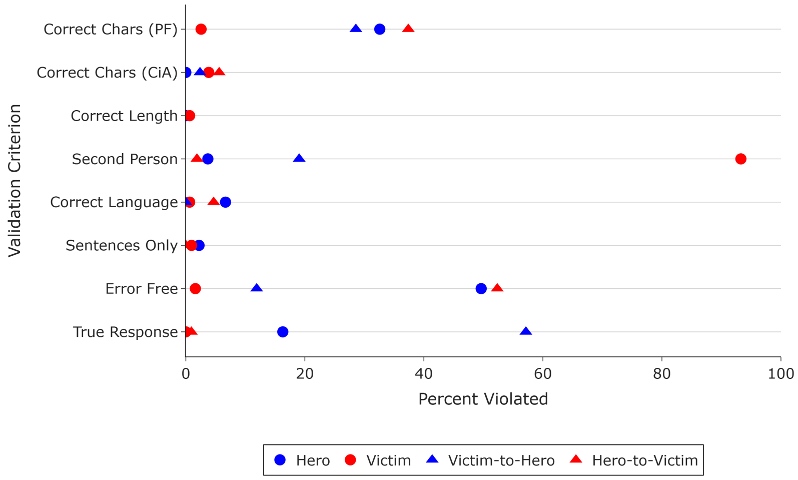

Figure 5.2: Violated Validation Criteria Across Message Type

Note: The criteria "On Topic" and "One or More Characters Used" were omitted because none of the invalid messages violated those criteria; "Correct Length" refers to the correct number of sentences.

Abbreviations: PF, Problem Framing; CiA, Characters in Action.

Word Overlap Between Constructed Messages

Hero and Victim characters used in messages had high overlap for both valid and invalid messages (Figure 5.3). Furthermore, there was less diversity of incorporated Hero characters in the messages compared to Victim characters (Figure 5.3). Hero messages incorporated significantly less unique Hero characters than initially given, reducing from the original 83 Hero characters to 34 Hero characters for valid messages and 29 Hero characters for invalid messages. Meanwhile, Victim messages incorporated almost all the 59 given Victim characters.

Sentiment Analysis on Constructed Messages

Individual message segments had very polarized sentiment across message types. Segments framed as Hero expressed very positive sentiments while segments framed as Victim expressed very negative sentiments (Figure 5.4). Expressed sentiment was slightly

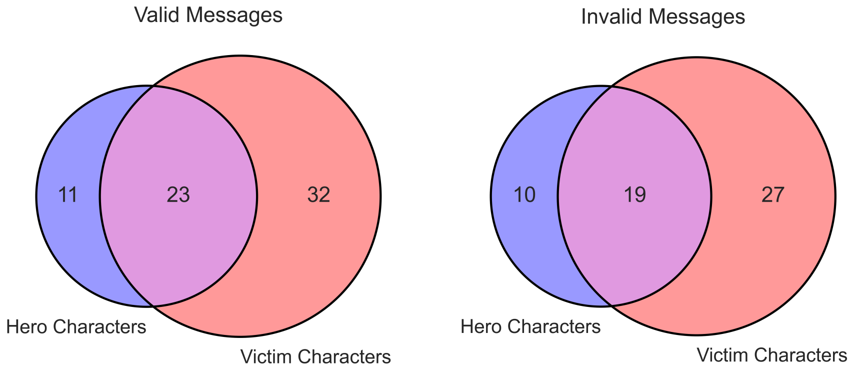

Figure 5.3: Character Overlap Across Valid and Invalid Messages

less polarized for Problem Framing segments compared to Characters in Action segments (Figure 5.4).

Whole messages skewed towards negative sentiments. Calculated sentiment ranged from "Negative" to "Very Negative" for all message types except for Hero, which had sentiments skew towards "Positive" (Figure 5.5). This skew was due to the Problem Definition and Scientific Information segments being classified as "Very Negative" and "Negative", respectively, which we attributed to the definitions provided by DHS CISA and other cybersecurity experts containing low-sentiment language.

Discussion & Future Work

Rapid Computational Risk Message Development

With respect to RQ #1, incorporating LLMs into risk communication message development enables fast message development and delivery. Our runtime analysis shows that *Llama* can speed up development time regardless of how the overall message is framed (Figure 5.1). Faster message development can lead to faster message deployment to affected populations [54], thus allowing people to have more time to prepare for and protect against

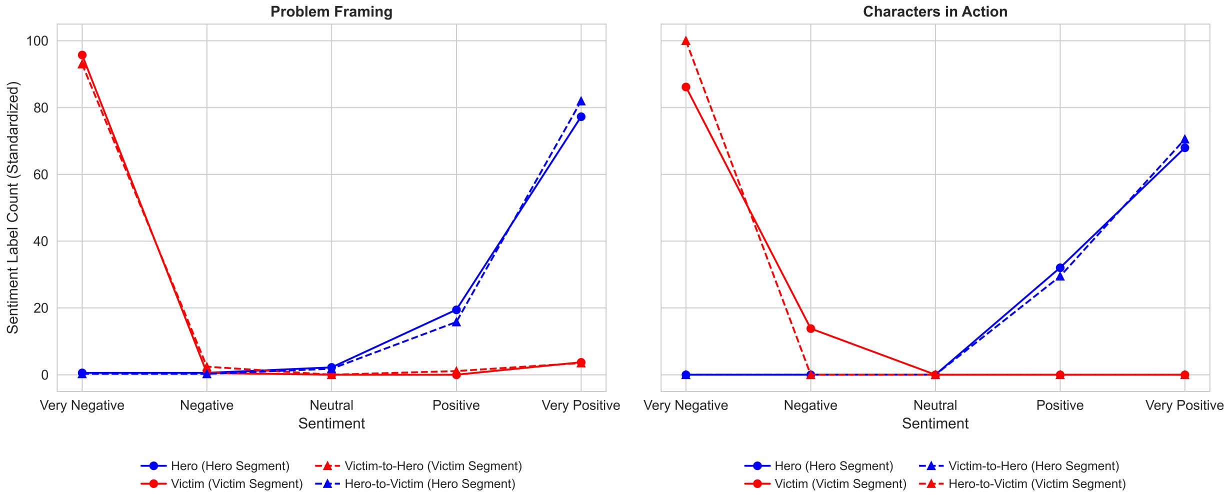

Figure 5.4: Sentiment of Llama-Generated Message Segments Across Message Type

a hazard like insider threats [82].

Rapid message development and deployment also relies on rapid validation. While we created a robust and thorough message validation criteria, our approach hinders the efficiency of development, a common problem in the risk communication research sphere [67, 82]. Human validation is a bottleneck. Yet, assessing LLM-generated segments in this manner provides more granular detail compared to most LLMs, given that LLMs lack human deductive reasoning skills [6] and can provide inaccurate information depending on their training data [67, 107]. Moreover, validating LLM-generated text with an LLM is inherently circular [96]. However, validation can be faster with the incorporation of various NLP techniques, such as sentiment analysis, frequency analysis, and topic modeling, to assist with validation steps while retaining human input and feedback [7, 68, 83]. Computationally enhanced validation, along with LLM-assisted message construction, is essential for fast insider threat risk communication development and deployment.

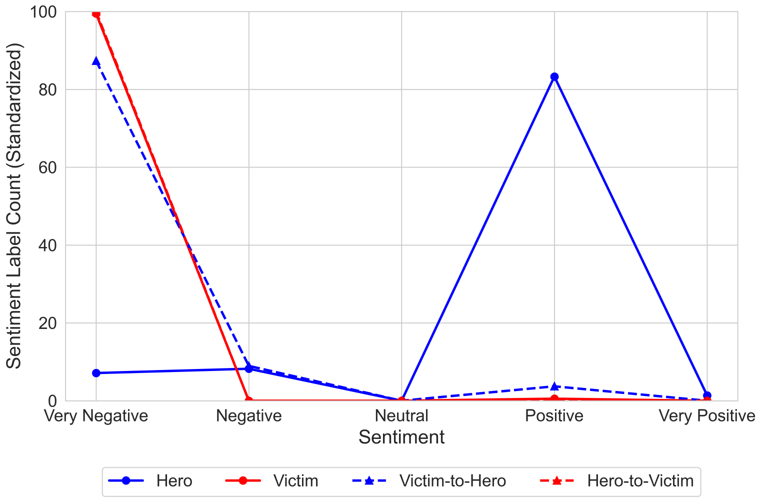

Figure 5.5: Sentiment of Insider Threat Risk Messages Across Message Type

Instrument Fidelity of Computational Risk Message Content

With respect to RQ #2, generating risk communication message segments with LLMs can produce precise, quality risk information and mitigation strategies. Our results show that *Llama* can generate valid risk messaging on insider threats when given clear instructions in tandem with zero-shot/few-shot learning examples (Tables 5.1 and 5.2). However, we found that Victim messages were generated with less precision and quality. Notably, *Llama* produced Victim messages that frequently violated the requirement that the text be written in the second person (Figure 5.2). Yet, all LLM messages still align with insider threats as the investigated hazard (see note under Figure 5.2). The fact that all messages created by *Llama* were both on topic and used appropriate character language indicates that our prompt engineering was well suited to *Llama*.

Implementing robust prompt engineering and hyperparameter tuning improves overall LLM performance [51]. LLM prompts engineered with explicit instructions and few-shot/zero-shot learning can produce consistent, accurate, and quality outputs [7, 51, 61, 68]. Additionally, LLM performance can be enhanced through fine-tuning model hyperparameters, incorporating human reinforcement learning, and providing training data specific to the needs of researchers [7, 8]. Our results here indicate that developing a specialized version of *Llama* for insider threat risk communication produces quality results with high precision (Table 5.2). Future risk communication research can benefit from training LLMs to be specialized for developing risk communication messages on other hazards.

Training an LLM, such as *Llama*, that can be run on an airgapped server is critical for security-sensitive applications. In previous research on risk communication for insider threats, the study of [68] had great success with cloud-based, proprietary LLMs. However, the risk of leaking sensitive information to adversaries is much greater on cloud based systems than running a local instance of an LLM on an airgapped computer [24, 47]. Given that AI engines, such as *ChatGPT*[14], are more mature than *Llama* [52], we are encouraged by the performance of *Llama* for developing insider threat risk messaging.

Instrument Fidelity of Computational Risk Message Structure

With respect to RQ #3, message segments generated by *Llama* generally adhere to the structure outlined by the NPF. These generated segments often frame insider threat messages accurately using characters from the provided character lists (Figures 5.2 and 5.3). However, character overlap can impact the validity of generated segments, which could explain why *Llama* sometimes frames messages with the wrong character type (Figure 5.2). This overlap often occurs when some characters can be classified as “Hero” in one context but as “Victim” in another [68, 94]. As an example, the character “critical infrastructure companies” was classified as both a Hero and Victim character. In Hero context, this character can be framed as an active participant in insider threat mitigation strategies whereas in Victim context, this character can be framed as an entity that suffers harm caused by insiders. Limiting the impacts character overlap can have on message framing requires prompt engineers to provide

[14] https://openai.com/

explicit few-shot learning examples in LLM training [7].

Our insider threat risk messages generally lean towards negative sentiments (Figure 5.5). However, message segments generated by *Llama* strongly align with one sentiment over the other depending on the character type specified. Our study shows that message segments explicitly framed as Hero lean towards positive sentiments whereas segments explicitly framed as Victim lean towards negative sentiments (Figure 5.4). Relationships between character framing in messages and sentiment are noted by NPF scholars, specifically the relationships between Hero messages and inducing positive affect in individuals [77, 83, 95]. Inducing a positive valence of affect through risk messaging can be important for motivating target populations towards preparedness [77, 82, 86]. Thus, future insider threat risk communication development may benefit from developing messages that elicit positive sentiment.

Future Directions with Computational Risk Message Development

Our work here continues the practical application of the DARC Framework by [68], covering steps aligning with message generation via LLMs [82]. We find that an LLM customized for risk communication development can improve message development efficiency without compromising the validity and reliability of resulting messages. However, some improvements to our process are needed to further improve study validity and reliability.

One aspect of our work to refine is additional *Llama* model fine-tuning. For future message development—for insider threats and beyond, we will incorporate robust hyperparameter tuning when customizing an LLM to improve model performance. Additionally, it would be beneficial to specify an adapter in our model, such as QLoRA [17], for low-cost LLM fine-tuning. Such methods would improve future risk communication development efficiency.

Another aspect to incorporate into our risk communication development pipeline is message efficacy testing. While the content analysis performed in [68] reveals message content and structure that can improve message efficacy, we cannot know for certain how effective these messages are without human subjects research. In the future, we will test our refined insider threat risk messages on a sample of individuals who could be exposed to insider threats in their field of work.

Threats to Validity

Construct Validity

We identified potential threats to construct validity [83] with our *Llama* model setup. These potential threats include model hallucination during response generation, vague prompt engineering, and infrequently updating a local base LLM instance. When fitting our *Llama* model, we implemented several measures to reduce as much response variability as possible. These measures included limiting the likelihood of model hallucination by lowering the model temperature [8], providing well-structured prompt information [61], and frequently updating our local base *Llama* model before applying our Modelfile.

Performing manual validation on messages also potentially threatens construct validity. The two message evaluators performed validation tasks without overlap. While validating messages with overlap would improve the reliability of our study, the research team did not have the resources to evaluate the large volume of generated messages with overlap. However, both evaluators were provided a stringent, well-defined message validation criteria to ensure similar assessment of message validity, and the validators also met to discuss individual messages when uncertain.

With respect to our use of the BERT model for sentiment analysis, we did not perform extensive model fine-tuning or hyperparameter optimization. Our application of BERT for message sentiment analysis was used to determine if a relationship between character type and sentiment was present, not to improve an already existing BERT model. Regardless, it is possible that optimizing the model could have produced refined sentiment analysis results specific to our risk communication development study.

Another threat to construct validity comes from our use of the insider threat character lists from the study of [68]. Characters identified as Hero and Victim in this study were determined using OpenAI's *GPT-4o mini* model[15], which can differ in performance based on how it is trained, developed, and optimized [52]. These character lists were repurposed—as opposed to re-identified with *Llama*—to keep the focus of the study on risk message generation steps. However, we are conducting ongoing research evaluating *Llama*'s performance at content analysis steps, which can be incorporated into future insider threat message development.

External Validity

External validity refers to how generalizable our study results are beyond its current context [67]. Similar to the study of [68], our focus on insider threats as the hazard and the NPF as the theoretical framework may be too specific to generalize to similar risk communication development or applied frameworks. However, the DARC Framework allows researchers to generalize risk communication development to any hazard, theoretical framework, or computational tools of choice [82].

Conclusion

LLMs aid efficient development of insider threat risk communication while promoting instrument fidelity. However, developing effective risk communication with LLMs still relies on human validation. This reliance on human input will likely decrease as LLMs and other computational analysis tools continue to advance.

[15] https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/

Computationally enhanced message development provides organizations a proactive and scalable approach for countering and mitigating insider threats. Proactively mitigating insider threats can minimize financial and infrastructural damage caused to organizations, which can lead them to be more resilient against future insider threats.

Acknowledgments

We thank Ryan Cummings and Kaveen Liyanage from the Montana State University (MSU) Software Engineering & Cybersecurity Lab (SECL) for their support and assistance with setting up a local instance of *Llama* on our Lambda Vector desktop system.

CONCLUSION

This thesis presents three manuscripts detailing the practical application of the DARC Framework with organizational insider threats as the hazard. The first manuscript compiles research on computational tools and theoretical frameworks spanning across multiple hazard domains. The second manuscript presents a mixed methods approach to content analysis guided by the DARC Framework, and the third manuscript continues this application for message construction steps.

The SLR presented in manuscript #1 goes in-depth about how past and current RCC research computationally develop messages. Out of the 26 articles reported on, the most common computational tools used are Natural Language Processing (NLP) techniques such as sentiment analysis and word classification, statistical modeling techniques such as logistic regression and cluster analysis, and content analysis software such as *NVivo*[1] or *Microsoft Excel*[2] (Table 3.5, Objective 1). Additionally, these articles discuss their application of social science frameworks to guide RCC development—the most commonly used frameworks being Protective Motivation Theory (PMT), the Crisis and Emergency Risk Communication (CERC) model, and the Narrative Policy Framework (NPF) (Table 3.6, Objective 2). The findings from this SLR provide researchers a compiled list of computational tools and theoretical frameworks to apply in future RCC development, and it also highlights critical research gaps present in RCC that can be addressed in the future.

One critical research gap uncovered in the SLR is the scant coverage of computational message construction. The most covered application of both computational tools and theoretical frameworks is for content analysis on hazard source text. While content analysis is an important prerequisite for constructing effective messages, more RCC research needs

[1] https://lumivero.com/products/nvivo/

[2] https://www.microsoft.com/en-us/microsoft-365/excel

to incorporate content analysis results into subsequent messaging. Hence the decision to integrate computational text analysis in *both* content analysis and subsequent message construction.

The work in manuscript #2 validates steps outlined by the DARC Framework that align with content analysis. I found that the integration of NLP and Large Language Models (LLMs) can sufficiently operationalize the NPF, uncovering message content and structure that can improve message efficacy (Objectives 3 and 4). This operationalization reveals that insider threat source text structured as the NPF Hero archetype can improve the efficacy of insider threat risk messaging (Figure 4.2), which I further investigate in manuscript #3. Determining effective message structure can be attributed to *GPT-4o-mini*'s ability to sufficiently replicate content coding using the NPF as a guide.

Nevertheless, I found that my LLM of choice, *GPT-4o-mini*, cannot fully replace human-centered content coding steps (Table 4.1, Objective 3). This is attributed to the LLM's inability at genuine deductive reasoning, resulting in less refined and nuanced responses compared to human RCC experts. Maintaining a human-assisted interaction between RCC experts and advanced computational text analysis tools is key for developing effective messaging.

This interaction between RCC experts and advanced computational tools continues in manuscript #3. This manuscript builds from manuscript #2 to construct effective risk communication on insider threats without compromising the validity and quality of the messages themselves. Based on the NLP and manual validation results, constructing insider threat messages with the LLM *Llama* efficiently constructs quality messaging adhering to the NPF while promoting instrument fidelity (Objectives 5 and 6). I attribute the model's solid performance to two factors: robust prompt engineering and model customization. With these two factors accounted for, future RCCs on insider threats can be constructed with very little deviation from given message construction tasks (Table 5.2).

Customizing an LLM for hazard- and RCC-specific tasks proves to be a contributing factor towards efficiently developing effective risk messages. However, as noted in the third manuscript, additional fine-tuning and testing measures are needed to further improve message construction instrument fidelity. Regarding extra fine-tuning, future LLM customization efforts should incorporate additional training datasets that are specific to not only the RCC hazard of choice but also to the desired RCC structure (e.g., NPF-structured messages). Training a custom LLM on RCC-specific data could improve the overall validity and quality of RCC messaging.

Regarding testing measures, while the work presented in manuscript #2 reveals message structure that could improve message efficacy, the only way to know whether subsequent messages are truly effective is through human subjects research. By testing these messages on a sample of individuals impacted by insider threats, I can gather feedback on how the messages are perceived and can be improved. Such feedback can be used to further refine insider threat RCCs to effectively reach as many affected individuals as possible.

Mitigating future harm from insider threats requires organizations to efficiently develop effective RCC messaging. Effective messaging encourages message recipients to adopt protective actions against insider threats. Concurrently, efficient message development leads to swift message delivery, providing affected employees more lead time to prepare for and mitigate harm caused by these types of threats. Through this thesis, I provide organizations, cybersecurity experts, and RCC researchers a comprehensive guide on insider threat mitigation through computationally enhanced RCC messaging. Such messaging will soften the financial and infrastructural damage caused by insider threats, creating more resilient and informed organizations across types and disciplines.

REFERENCES CITED

[1] Elissa M Abrams, Marcus Shaker, and Matthew Greenhawt. Covid-19 and the importance of effective risk communication with children. *Paediatrics Child Health*, 27(Supplement_1):S1–S3, 2022.

[2] Ghadah Adel and Yuping Wang. Arabic twitter corpus for crisis response messages classification. In *Proceedings of the 2019 2nd International Conference on Algorithms, Computing and Artificial Intelligence*, pages 498–503–498–503. Association for Computing Machinery.

[3] Yusuf Albayram, John Liu, and Stivi Cangonj. Comparing the effectiveness of text-based and video-based delivery in motivating users to adopt a password manager. In *Proceedings of the 2021 European Symposium on Usable Security*, pages 89–104–89–104. Association for Computing Machinery.

[4] Fatimah Mohammed Alhassan and Sharifah Abdullah AlDossary. The saudi ministry of health's twitter communication strategies and public engagement during the covid-19 pandemic: Content analysis study. *JMIR Public Health Surveill*, 7(7):e27942–e27942, 2021. 2369-2960.

[5] Fatima Rashed Alzaabi and Abid Mehmood. A review of recent advances, challenges, and opportunities in malicious insider threat detection using machine learning methods. *IEEE Access*, 12:30907–30927, 2024.

[6] Maryam Amirizaniani, Elias Martin, Maryna Sivachenko, Afra Mashhadi, and Chirag Shah. Can llms reason like humans? assessing theory of mind reasoning in llms for open-ended questions. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, page 34–44, New York, NY, USA, 2024. Association for Computing Machinery.

[7] D.M. Anisuzzaman, Jeffrey G. Malins, Paul A. Friedman, and Zachi I. Attia. Fine-tuning large language models for specialized use cases. *Mayo Clinic Proceedings: Digital Health*, 3(1):100184, 2025.

[8] Liam Barkley and Brink van der Merwe. Investigating the role of prompting and external tools in hallucination rates of large language models, 2024.

[9] A. Bartolucci, M. C. Aquilino, L. Bril, J. Duncan, and T. van Steen. Effectiveness of audience segmentation in instructional risk communication: A systematic literature review. *International Journal of Disaster Risk Reduction*, 95:103872, 2023.

[10] Victor R. Basili, Gianluigi Caldiera, and Hans Dieter Rombach. The goal question metric approach. 1994.

[11] Scott R. Boss, Dennis F. Galletta, Paul Benjamin Lowry, Gregory D. Moody, and Peter Polak. What do systems users have to fear? using fear appeals to engender threats and fear that motivate protective security behaviors. *MIS Quarterly*, 39(4):837–864, 2015.

[12] Alan Bryman. *Social research methods.* Oxford university press, 2016.

[13] Bowen Cao, Deng Cai, Zhisong Zhang, Yuexian Zou, and Wai Lam. On the worst prompt performance of large language models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 69022–69042. Curran Associates, Inc., 2024.

[14] Tianying Chen, Margot Stewart, Zhiyu Bai, Eileen Chen, Laura Dabbish, and Jessica Hammer. Hacked time: Design and evaluation of a self-efficacy based cybersecurity game. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference*, pages 1737–1749–1737–1749. Association for Computing Machinery.

[15] Michael F. Dahlstrom. Using narratives and storytelling to communicate science with nonexpert audiences. *Proceedings of the National Academy of Sciences*, 111(supplement_4):13614–13620, 2014.

[16] Nic DePaula, Loni Hagen, Stiven Roytman, and Dana Alnahass. Platform effects on public health communication: A comparative and national study of message design and audience engagement across twitter and facebook. *JMIR Infodemiology*, 2(2):e40198–e40198, 2022. 2564-1891.

[17] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms, 2023.

[18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the NAACL HLT*, pages 4171–4186. Association for Computational Linguistics, 2019.

[19] Larissa S. Drescher, Jutta Roosen, Katja Aue, Kerstin Dressel, Wiebke Schär, and Anne Götz. Emotionality in covid-19 crisis communication from authorities and independent experts on twitter. *Federal Health Gazette - Health Research - Health Protection*, 66:689–699, 2023.

[20] Ali Farghaly and Khaled Shaalan. Arabic natural language processing: Challenges and solutions. *ACM Transactions on Asian Language Information Processing*, 8(4):Article 14, 2009.

[21] A. Fathollahzadeh, I. Salmani, M. A. Morowatisharifabad, M. R. Khajehaminian, J. Babaie, and H. Fallahzadeh. Models and components in disaster risk communication: A systematic literature review. (2277-9531 (Print)), 2023.

[22] Stefano Filippi. Measuring the impact of chatgpt on fostering concept generation in innovative product design. *Electronics*, 12(16):3535, 2023.

[23] Baruch Fischhoff and Julie S. Downs. Communicating foodborne disease risk. *Emerging Infectious Diseases*, 3(4):489–495, 1997.

[24] Othmane Friha, Mohamed Amine Ferrag, Burak Kantarci, Burak Cakmak, Arda Ozgun, and Nassira Ghoualmi-Zine. Llm-based edge intelligence: A comprehensive survey on architectures, applications, security and trustworthiness. *IEEE Open Journal of the Communications Society*, 5:5799–5856, 2024.

[25] Abdul Ghafoor, Ali Shariq Imran, Sher Muhammad Daudpota, Zenuun Kastrati, Abdullah, Rakhi Batra, and Mudasir Ahmad Wani. The impact of translating resource-rich datasets to low-resource languages through multi-lingual text processing. *IEEE Access*, 9:124478–124490, 2021.

[26] Stephen Gilbert, Hugh Harvey, Tom Melvin, Erik Vollebregt, and Paul Wicks. Large language model ai chatbots require approval as medical devices. *Nature Medicine*, 29(10):2396–2398–2396–2398, 2023.

[27] Ross Gore, Ann Marie Reinhold, Barry Ezell, Christopher J. Lynch, Clemente Izurieta, Jessica O'Brien, Virginia Zamponi, Madison Munro, Erik Jensen, and Elizabeth A. Shanahan. Building a domain agnostic framework for efficient and effective risk communication messages. In *MODSIM World 2024*, pages 1–11. MODSIM World.

[28] Melanie C Green and Timothy C Brock. The role of transportation in the persuasiveness of public narratives. *Journal of personality and social psychology*, 79(5):701, 2000.

[29] Frank Greitzer, Justin Purl, Yung Mei Leong, and D.E. Sunny Becker. Sofit: Sociotechnical and organizational factors for insider threat. In *2018 IEEE Security and Privacy Workshops (SPW)*, pages 197–206, 2018.

[30] Frank L. Greitzer, Jeremy Strozer, Sholom Cohen, John Bergey, Jennifer Cowley, Andrew Moore, and David Mundie. Unintentional insider threat: Contributing factors, observables, and mitigation strategies. In *2014 47th Hawaii International Conference on System Sciences*, pages 2025–2034, 2014.

[31] Sara K. Guenther and Elizabeth A. Shanahan. Communicating risk in human-wildlife interactions: How stories and images move minds. *PLOS ONE*, 15(12):e0244440, 2021.

[32] Timothy C Guetterman, Tammy Chang, Melissa DeJonckheere, Tanmay Basu, Elizabeth Scruggs, and VG Vinod Vydiswaran. Augmenting qualitative text analysis with natural language processing: Methodological study. *J Med Internet Res*, 20(6):e231, 2018.

[33] Thilo Hagendorff and David Danks. Ethical and methodological challenges in building morally informed ai systems. *AI and Ethics*, 3(2):553–566, 2023.

[34] Karin Hannes and Pieter Thyssen. Towards an inclusive covid-19 crisis communication policy in belgium: the development and validation of strategies for multilingual and media accessible crisis communication. deliverable 1: Scientific evidence feeding into the guideline development process - rapid systematic literature review. Report, Sciensano, 2022.

[35] Mary Clare Hano, Steven E. Prince, Linda Wei, Bryan J. Hubbell, and Ana G. Rappold. Knowing your audience: A typology of smoke sense participants to inform wildfire smoke health risk communication. *Frontiers in Public Health*, 8:2296–2565, 2020.

[36] Marian Harbach, Markus Hettig, Susanne Weber, and Matthew Smith. Using personal examples to improve risk communication for security & privacy decisions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, page 2647–2656, New York, NY, USA, 2014. Association for Computing Machinery.

[37] Robert L. Heath, Jaesub Lee, Michael J. Palenchar, and Laura L. Lemon. Risk communication emergency response preparedness: Contextual assessment of the protective action decision model. *Risk Analysis*, 38(2):333–344, 2018.

[38] Thomas F. Heston and Charya Khun. Prompt engineering in medical education. *International Medical Education*, 2(3):198–205, 2023.

[39] Michael D. Jones, Elizabeth A. Shanahan, and Mark K. McBeth. *The science of stories: Applications of the narrative policy framework in public policy analysis*. Springer, 2014.

[40] Youngkee Ju and Myoungsoon You. Developing a gist-extraction typology based on journalistic lead writing: A case of food risk news. *Heliyon*, 4(8):e00738–e00738, 2018. 2405-8440.

[41] Katikapalli Subramanyam Kalyan. A survey of gpt-3 family large language models including chatgpt and gpt-4. *Natural Language Processing Journal*, 6:100048, 2024.

[42] Mert Karabacak and Konstantinos Margetis. Embracing large language models for medical applications: Opportunities and challenges. *CUREUS JOURNAL OF MEDICAL SCIENCE*, 15(5), MAY 21 2023.

[43] Elise Karinshak, Sunny Xun Liu, Joon Sung Park, and Jeffrey T. Hancock. Working with ai to persuade: Examining a large language model's ability to generate pro-vaccination messages. *Proc. ACM Hum.-Comput. Interact.*, 7(CSCW1):Article 116, 2023.

[44] Georges Elias Khalil, Karen S. Calabro, Brittani Crook, Tamara C. Machado, Cheryl L. Perry, and Alexander V. Prokhorov. Validation of mobile phone text messages for nicotine and tobacco risk communication among college students: A content analysis. *Tobacco Prevention Cessation*, 4(February), 2018.

[45] Sebastian Krügel, Andreas Ostermaier, and Matthias Uhl. Chatgpt's inconsistent moral advice influences users' judgment. *Scientific Reports*, 13(1):4569, 2023.

[46] Andrei Kucharavy. *From Deep Neural Language Models to LLMs*, pages 3–17. Springer Nature Switzerland, Cham, 2024.

[47] B. V. Pranay Kumar and M. D. Shaheer Ahmed. Beyond clouds: Locally runnable llms as a secure solution for ai applications. *Digital Society*, 3(49), 2024.

[48] Raul P. Lejano, Eulito V. Casas, Rosabella B. Montes, and Lynie P. Lengwa. Weather, climate, and narrative: A relational model for democratizing risk communication. *Weather, Climate, and Society*, 10(3):579–594, 2018.

[49] Miłosz Lewandowski, Paweł Łukowicz, Dariusz Świetlik, and Wioletta Barańska-Rybak. Chatgpt-3.5 and chatgpt-4 dermatological knowledge level based on the specialty certificate examination in dermatology. *Clinical and Experimental Dermatology*, 49(7):686–691, 08 2023.

[50] Jundong Li, Kewei Cheng, Dakuo Wang, Fred Morstatter, Robert P. Trevino, Jiliang Tang, and Huan Liu. Feature selection: A data perspective. *ACM Comput. Surv.*, 50(6):Article 94, 2017.

[51] Sue Lim and Ralf Schmälzle. Artificial intelligence for health message generation: an empirical study using a large language model (llm) and prompt engineering. *Frontiers in Communication*, 8, 2023.

[52] Trevor Lin, Ryan T. Lin, Rahul Mhaskar, and Curtis E. Margo. Evaluating the accuracy of advanced language learning models in ophthalmology: A comparative study of chatgpt-4o and meta ai's llama 3.1. *Advances in ophthalmology practice and research*, 5(2):95–99, 2025.

[53] Liu Liu, Olivier De Vel, Qing-Long Han, Jun Zhang, and Yang Xiang. Detecting and preventing cyber insider threats: A survey. *IEEE Communications Surveys Tutorials*, 20(2):1397–1417, 2018.

[54] May O. Lwin, Jiahui Lu, Anita Sheldenkar, and Peter J. Schulz. Strategic uses of facebook in zika outbreak communication: Implications for the crisis and emergency risk communication model. *International Journal of Environmental Research and Public Health*, 15(9):1974, 2018.

[55] Christopher J Lynch, Erik Jensen, Madison H Munro, Virginia Zamponi, Joseph Martinez, Kevin O'Brien, Brandon Feldhaus, Katherine Smith, Ann Marie Reinhold, and Ross Gore. Gpt-4 generated narratives of life events using a structured narrative prompt: A validation study. *arXiv preprint arXiv:2402.05435*, page 29, 2024.

[56] Christopher J. Lynch, Erik J. Jensen, Virginia Zamponi, Kevin O'Brien, Erika Frydenlund, and Ross Gore. A structured narrative prompt for prompting narratives from large language models: Sentiment assessment of chatgpt-generated narratives and real tweets. *Future Internet*, 15(12):375, 2023.

[57] Melissa MacKay, Andrea Cimino, Samira Yousefinaghani, Jennifer E. McWhirter, Rozita Dara, and Andrew Papadopoulos. Canadian covid-19 crisis communication on twitter: Mixed methods research examining tweets from government, politicians, and public health for crisis communication guiding principles and tweet engagement. *International Journal of Environmental Research and Public Health*, 19(11):1660–4601, 2022.

[58] Melissa MacKay, Taylor Colangeli, Daniel Gillis, Jennifer McWhirter, and Andrew Papadopoulos. Examining social media crisis communication during early covid-19 from public health and news media for quality, content, and corresponding public sentiment. *International Journal of Environmental Research and Public Health*, 18(15):1660–4601, 2021.

[59] Melissa MacKay, Caitlin Ford, Taylor Colangeli, Daniel Gillis, Jennifer E. McWhirter, and Andrew Papadopoulos. A content analysis of canadian influencer crisis messages on instagram and the public's response during covid-19. *BMC Public Health*, 22(1):763–763, 2022.

[60] James E. Maddux and Ronald W. Rogers. Protection motivation and self-efficacy: A revised theory of fear appeals and attitude change. *Journal of Experimental Social Psychology*, 19(5):469–479, 1983.

[61] Seyed Kourosh Mahjour, Ramin Soltanmohammadi, Ehsan Heidaryan, and Salah A. Faroughi. Geosystems risk and uncertainty: The application of chatgpt with targeted prompting. *Geoenergy Science and Engineering*, 238:212889, 2024.

[62] R. Manimegalai, S. Kavisri, M. Vasundhra, and R. Kingsy Grace. Machine learning framework for analyzing disaster-tweets. In *2023 International Conference on Intelligent Systems for Communication, IoT and Security (ICISCoIS)*, pages 55–60.

[63] Yanying Mao, Qun Liu, and Yu Zhang. Sentiment analysis methods, applications, and challenges: A systematic literature review. *Journal of King Saud University - Computer and Information Sciences*, 36(4):102048, 2024.

[64] Mary L. McHugh. Interrater reliability: the kappa statistic. *Biochemia Medica*, 22(3):276–282, 2012.

[65] Bertalan Meskó. The impact of multimodal large language models on health care's future. *J Med Internet Res*, 25:e52865, 2023.

[66] Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M. Krumholz, Jure Leskovec, Eric J. Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, 2023.

[67] Madison H. Munro, Ross J. Gore, Christopher J. Lynch, Yvette D. Hastings, and Ann Marie Reinhold. Enhancing risk and crisis communication with computational methods: A systematic literature review. *Risk Analysis*, 45(7):1683–1697, 2025.

[68] Madison H. Munro, Manuel Ruiz-Aravena, Elizabeth A. Shanahan, Savanna Washburn, and Ann Marie Reinhold. Integrating computational text analysis into risk and crisis communication development. In *2025 IEEE International Conference on Information Reuse and Integration and Data Science (IRI)*, pages 373–378, 2025.

[69] A. H. Nasution and A. Onan. Chatgpt label: Comparing the quality of human-generated and llm-generated annotations in low-resource language nlp tasks. *IEEE Access*, 12:71876–71900, 2024.

[70] Laura K. Nelson. Computational grounded theory: A methodological framework. *Sociological Methods Research*, 49(1):3–42, 2020.

[71] Laura K. Nelson, Derek Burk, Marcel Knudsen, and Leslie McCall. The future of coding: A comparison of hand-coding and three types of computer-assisted text analysis methods. *Sociological Methods & Research*, 50(1):202–237, 2021.

[72] R. I. Ogie, J. C. Rho, and R. J. Clarke. Artificial intelligence in disaster risk communication: A systematic literature review. In *2018 5th International Conference on Information and Communication Technologies for Disaster Management (ICT-DM)*, pages 1–8.

[73] Michele K. Olson, Jeannette Sutton, Sarah C. Vos, Robert Prestley, Scott L. Renshaw, and Carter T. Butts. Build community before the storm: The national weather service's social media engagement. *Journal of Contingencies and Crisis Management*, 27(4):359–373, 2019.

[74] Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, Roger Chou, Julie Glanville, Jeremy M Grimshaw, Asbjørn Hróbjartsson, Manoj M Lalu, Tianjing Li, Elizabeth W Loder, Evan Mayo-Wilson, Steve McDonald, Luke A McGuinness, Lesley A Stewart, James Thomas, Andrea C Tricco, Vivian A Welch, Penny Whiting, and David Moher. The prisma 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, 372:n71, 2021.

[75] Matthew J Page, David Moher, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, Roger Chou, Julie Glanville, Jeremy M Grimshaw, Asbjørn Hróbjartsson, Manoj M Lalu, Tianjing Li, Elizabeth W Loder, Evan Mayo-Wilson, Steve McDonald, Luke A McGuinness, Lesley A Stewart, James Thomas, Andrea C Tricco, Vivian A Welch, Penny Whiting, and Joanne E McKenzie. Prisma 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. *BMJ*, 372:n160, 2021.

[76] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2024.

[77] Eric D. Raile, Elizabeth A. Shanahan, Richard C. Ready, Jamie McEvoy, Clemente Izurieta, Ann Marie Reinhold, Geoffrey C. Poole, Nicolas T. Bergmann, and Henry King. Narrative risk communication as a lingua franca for environmental hazard preparation. *Environmental Communication*, 16(1):108–124, 2022.

[78] Sriram Ramanathan, Lisa-Angelique Lim, Nazanin Rezazadeh Mottaghi, and Simon Buckingham Shum. When the prompt becomes the codebook: Grounded prompt engineering (groproe) and its application to belonging analytics. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, LAK '25, page 713–725, New York, NY, USA, 2025. Association for Computing Machinery.

[79] Usman Rauf, Fadi Mohsen, and Zhiyuan Wei. A taxonomic classification of insider threats: Existing techniques, future directions recommendations. *Journal of Cyber Security and Mobility*, 12(2):221–252, 2023.

[80] Gabriel Recchia, Alice C. E. Lawrence, and Alexandra L. J. Freeman. Investigating the presentation of uncertainty in an icon array: A randomized trial. *PEC Innovation*, 1:100003–100003, 2022. 2772-6282.

[81] Ann Marie Reinhold, Ross J. Gore, Barry Ezell, Clemente I. Izurieta, and Elizabeth A. Shanahan. From cyclones to cybersecurity: A call for convergence in risk and crisis communications research. *Journal of Homeland Security and Emergency Management*, 2025.

[82] Ann Marie Reinhold, Madison H. Munro, Elizabeth A. Shanahan, Ross J. Gore, Barry C. Ezell, and Clemente Izurieta. Embedding software engineering in mixed methods: Computationally enhanced risk communication. *IJMRA*, 15:67–72, 2023.

[83] Ann Marie Reinhold, Eric D. Raile, Clemente Izurieta, Jamie McEvoy, Henry W. King, Geoffrey C. Poole, Richard C. Ready, Nicolas T. Bergmann, and Elizabeth A. Shanahan. Persuasion with precision: Using natural language processing to improve instrument fidelity for risk communication experimental treatments. *Journal of Mixed Methods Research*, 17(4):373–395, 2023.

[84] Barbara Reynolds and Matthew W. Seeger. Crisis and emergency risk communication as an integrative model. *Journal of Health Communication*, 10(1):43–55, 2005. doi: 10.1080/10810730590904571.

[85] Charis Rice and Rosalind H. Searle. ‘the enabling role of internal organizational communication in insider threat activity – evidence from a high security organization’. *Management Communication Quarterly*, 36(3):467–495, 2022.

[86] Laura N. Rickard. Pragmatic and (or) constitutive? on the foundations of contemporary risk communication research. *Risk Analysis*, 41(3):466–479, 2021.

[87] Heather Riddell and Christopher Fenner. User-generated crisis communication: Exploring crisis frames on twitter during hurricane harvey. *Southern Communication Journal*, 86(1):31–45, 2021.

[88] Malik Sallam. Chatgpt utility in healthcare education, research, and practice: Systematic review on the promising perspectives and valid concerns, 2023.

[89] Ruta Sawant and Sujit Sansgiry. Communicating risk of medication side-effects: role of communication format on risk perception. *Pharmacy Practice (Granada)*, 16(2), 2018.

[90] Joshua Schimel. *Writing Science: How to Write Papers That Get Cited and Proposals That Get Funded*. Oxford University Press, 2012.

[91] Timothy L Sellnow, Robert R Ulmer, Matthew W Seeger, and Robert Littlefield. *Effective risk communication: A message-centered approach*. Food Microbiology and Food Safety. Springer Science Business Media, New York, 2008.

[92] S. Selva Birunda and R. Kanniga Devi. A review on word embedding techniques for text classification. In Jennifer S. Raj, Abdullah M. Iliyasu, Robert Bestak, and Zubair A. Baig, editors, *Innovative Data Communication Technologies and Application*, pages 267–281, Singapore, 2021. Springer Singapore.

[93] Elizabeth A. Shanahan, Rob A. DeLeo, Elizabeth A. Albright, Meng Li, Elizabeth A. Koebele, Kristin Taylor, Deserai Anderson Crow, Katherine L. Dickinson, Honey Minkowitz, Thomas A. Birkland, and Manli Zhang. Visual policy narrative messaging improves covid-19 vaccine uptake. *PNAS Nexus*, 2(4):pgad080, 2023.

[94] Elizabeth A. Shanahan, Michael D. Jones, Mark K. McBeth, and Claudio M. Radaelli. *The narrative policy framework*, page 173–213. Westview Press, 4th edition, 2018.

[95] Elizabeth A. Shanahan, Ann Marie Reinhold, Eric D. Raile, Geoffrey C. Poole, Richard C. Ready, Clemente Izurieta, Jamie McEvoy, Nicolas T. Bergmann, and Henry King. Characters matter: How narratives shape affective responses to risk communication. *PLOS ONE*, 14(12):1–24, 2019.

[96] Shreya Shankar, J.D. Zamfirescu-Pereira, Bjoern Hartmann, Aditya Parameswaran, and Ian Arawjo. Who validates the validators? aligning llm-assisted evaluation of llm outputs with human preferences. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, UIST '24, New York, NY, USA, 2024. Association for Computing Machinery.

[97] Prativa Sharma, Bandana Kar, Jun Wang, and Doug Bausch. A machine learning approach to flood severity classification and alerting. In *Proceedings of the 4th ACM SIGSPATIAL International Workshop on Advances in Resilient and Intelligent Cities*, pages 42–47–42–47. Association for Computing Machinery.

[98] Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context, 2023.

[99] Max Silberztein. *The Limitations of Corpus-Based Methods in NLP*, pages 3–24. Springer Nature Switzerland, Cham, 2024.

[100] Julia Silge and David Robinson. tidytext: Text mining and analysis using tidy data principles in r. *JOSS*, 1(3), 2016.

[101] Catherine E. Slavik, Charlotte Buttle, Shelby L. Sturrock, J. Connor Darlington, and Niko Yiannakoulias. Examining tweet content and engagement of canadian public health agencies and decision makers during covid-19: Mixed methods analysis. *J Med Internet Res*, 23(3):e24883–e24883, 2021. 1438-8871.

[102] Joanna Sleigh, Julia Amann, Manuel Schneider, and Effy Vayena. Qualitative analysis of visual risk communication on twitter during the covid-19 pandemic. *BMC public health*, 21:1–12–1–12, 2021.

[103] Sonia H. Stephens and Daniel P. Richards. Story mapping and sea level rise: listening to global risks at street level. *Commun. Des. Q. Rev*, 8(1):5–18–5–18, 2020.

[104] Chris Stokel-Walker and Richard van Noorden. What chatgpt and generative ai mean for science. *Nature*, 614:214–216, 2023.

[105] Omri Suissa, Avshalom Elmalech, and Maayan Zhitomirsky-Geffet. Text analysis using deep neural networks in digital humanities and information science. *Journal of the Association for Information Science and Technology*, 73(2):268–287, 2022.

[106] Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. Large language models in medicine. *Nature medicine*, 29(8):1930–1940–1930–1940, 2023.

[107] Eva A. M. van Dis, Johan Bollen, Willem Zuidema, Robert van Rooij, and Claudi L. Bockting. Chatgpt: Five priorities for research. *Nature*, 2023.

[108] Sarah C. Vos, Jeannette Sutton, Yue Yu, Scott Leo Renshaw, Michele K. Olson, C. Ben Gibson, and Carter T. Butts. Retweeting risk communication: The role of threat and efficacy. *Risk Analysis*, 38(12):2580–2598, 2018.

[109] Dakuo Wang, Justin D. Weisz, Michael Muller, Parikshit Ram, Werner Geyer, Casey Dugan, Yla Tausczik, Horst Samulowitz, and Alexander Gray. Human-ai collaboration in data science: Exploring data scientists' perceptions of automated ai. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW):Article 211, 2019.

[110] Jing Wang, Huan Deng, Bangtao Liu, Anbin Hu, Jun Liang, Lingye Fan, Xu Zheng, Tong Wang, and Jianbo Lei. Systematic evaluation of research progress on natural language processing in medicine over the past 20 years: Bibliometric study on pubmed. *Journal of Medical Internet Research*, 22(1), 2020.

[111] Hyejung Yoon, Myoungsoon You, and Changwoo Shon. An application of the extended parallel process model to protective behaviors against covid-19 in south korea. *PLOS ONE*, 17(3):1–15, 2022.

[112] Haoran Zhang, Amy X. Lu, Mohamed Abdalla, Matthew McDermott, and Marzyeh Ghassemi. Hurtful words: quantifying biases in clinical contextual word embeddings, 2020.

[113] Xiaochen Angela Zhang. Understanding the cultural orientations of fear appeal variables: a cross-cultural comparison of pandemic risk perceptions, efficacy perceptions, and behaviors. *Journal of Risk Research*, 24(3-4):432–448, 2021. doi: 10.1080/13669877.2021.1887326.

[114] Xiaochen Angela Zhang and Jonathan Borden. How to communicate cyber-risk? an examination of behavioral recommendations in cybersecurity crises. *Journal of Risk Research*, 23(10):1336–1352, 2020.

[115] Xueying Zhang and Shuhua Zhou. Sharing health risk messages on social media: Effects of fear appeal message and image promotion. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, 14(2), 2020.

[116] Xinyan Zhao, Mengqi Monica Zhan, and Brooke Fisher Liu. Understanding motivated publics during disasters: Examining message functions, frames, and styles of social media influentials and followers. *Journal of Contingencies and Crisis Management*, 27(4):387–399, 2019.

[117] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc., 2023.