Towards Automatic Identification of Task-Based Cyber Risk

Corey Bolger[1][0009−0009−5284−1614], Raghavendra Cherupalli[1], Samantha Phillips[2][0009−0004−8960−4966], Bradley Brummel[3][0000−0003−1400−2034], Sal Aurigemma[1][0000−0002−2057−9326], and Tyler Moore[1][0000−0002−8771−8191]

[1] University of Tulsa, Tulsa OK 74104, USA
[2] Kennesaw State University, Kennesaw GA 30144, USA
[3] University of Houston, Houston Texas 77204, USA

Abstract. Users are often flagged as an important source of cyber risk. However, in enterprise environments, some job tasks are riskier than others. We construct a working definition of task-based cyber risk and apply it to tasks from the O*NET database of occupation profiles. We first conduct a qualitative analysis using human coders, then develop and evaluate an automated approach using large language models. The results demonstrate the feasibility of identifying cyber risk from job task descriptions, which could be used to manage enterprise cyber risk better.

Keywords: Cyber risk · Large Language Model · Job tasks.

1 Introduction

Understanding cybersecurity risk is crucial for any organization aiming to improve its security. Cyber risk management programs guide security-related decisions, relying on information provided by industry partners, internal security teams, and government advisories. A common refrain in these analyses is the practice of treating users as a significant risk factor [35, 29, 13]. This viewpoint has received pushback as being misguided victim-blaming [1].

An alternative approach could be to focus on the specific tasks users undertake that exhibit elevated cyber risk. This approach contributes in at least two meaningful ways. First, the focus on tasks shifts responsibility from the individual to the job role and its place in the organization. Second, identifying tasks associated with elevated cyber risk creates opportunities for mitigation through training, process change, or the adoption of controls.

Specific tasks that users perform rarely appear in current cyber risk analyses. Where distinctions are made, users are typically categorized as regular or those with administrative privileges, with executives and other VIPs sometimes placed into a separate category for monitoring. The approach presented in this paper seeks to be more precise in delineating tasks and the cyber risks they exhibit. We devise an automated method to label job tasks that contain cyber risk. The method utilizes a custom task-based cyber risk definition, derived from iterative

grounded theory, to determine whether specific tasks sourced from a large jobs database have cyber risk. Categorizing tasks based on the presence of cyber risk could enable security practitioners to better determine whether additional security controls are needed, whether users require task-specific security training, or whether the task should be redesigned to be performed securely.

2 Research Vision for Task-Based Cyber Risk

This paper is narrowly focused on labeling job tasks for the presence of cyber risk. Ultimately, our aim is to apply this method to large, comprehensive databases of job roles and associated tasks. We briefly discuss potential avenues for future research to help motivate the objectives of the present paper.

Tailoring Cyber Risk to Work Roles. Organizations that have their own internal job and task descriptions could apply this method to construct a database of task-based cyber risk specific to their own environment. This information in turn could be used to better manage cybersecurity risk facing the organization. Targeted training could be designed for risky tasks and delivered to just those affected workers. By focusing on actual job-related tasks, the training is also inherently more relevant than general-purpose training. Moreover, workers identified as engaging in riskier tasks could be supported through enhanced security monitoring and protective controls. Finally, the organization could choose to alter or eliminate tasks deemed to be too high-risk.

Such tailoring could also be utilized beyond individual organizations, since many job roles are common across companies. Once again, training and security controls could be adapted to the identified tasks present for particular roles, creating a more granular understanding of cyber risk and how it can be mitigated.

Evaluating Organizational-Level Cyber Risk. At the organization level, this research enables a broader, more strategic view of cyber risk by aggregating task-based insights across roles. Rather than focusing on individual users or tasks in isolation, this perspective allows organizations to evaluate how cyber risk is distributed and managed across the enterprise.

Organizations could conduct gap assessments of existing security controls relative to identified task-based risks. This can reveal potential weaknesses in current controls and highlight areas where additional cybersecurity investment may be most effective. New or proposed controls can also be evaluated against these identified risks to determine whether they adequately mitigate them.

Additionally, task-based cyber risk data could be mapped to known attack vectors from frameworks such as MITRE ATT&CK [26]. This would enable a better understanding of not only the attack techniques themselves but also the organizational conditions that make those attacks more likely. Additionally, this approach could support the development of organization-level risk scores based on the roles and tasks present within a given organization. Such scoring mechanisms could be used by external stakeholders, such as cyber insurance

providers, to estimate organizational risk exposure based on the types of cyber risks employees are likely to encounter.

3 Background

This research draws on multiple domains to establish a foundation for task-based cyber risk identification. It begins with enterprise risk management, then examines the sources and characteristics of occupational data as used in industrial-organizational psychology, and finally considers the application of large language models to automate the classification of large datasets.

Cybersecurity Risks Both academic researchers and industry practitioners have developed ways to measure enterprise risk [10, 27]. The National Institute of Standards and Technology has published a Risk Management Framework (RMF) designed to help organizations manage security and privacy risk. Organizations can use this information to determine their security investments [15, 16].

Human aspects of cybersecurity risk have been widely researched. Studies have shown the extent to which organizations go to modify human behavior to reduce cybersecurity risk [5]. Tasks known to have inherent cyber risk, such as sending and receiving email, have received specific attention [22, 23, 33]. Furthermore, research has also indicated that work tasks can come into conflict with security policies, resulting in insecure user behavior [19]. Recent work has explored the possibility of assigning users different cyber risk profiles based upon their role [6]. Despite this, there has been little research connecting cybersecurity training to specific job tasks [28].

Occupation Research Industrial-organizational psychologists have developed a range of methods for analyzing jobs and the tasks within those jobs. They have applied the methods across a wide range of industries and occupations [2, 32]. Significant research has also been conducted around the concepts of job design and job crafting. These concepts refer to both the top-down structuring of jobs by organizations as well as the bottom-up modification of jobs by employees [31, 21]. The goal of these processes is to improve employee performance within roles.

Brummel et al. applied these principles to organizational cybersecurity by categorizing employees into one of four different roles, users, administrators, engineers, and executives [7]. They identified several key tasks related to cybersecurity and highlighted that "categorizing employees by their roles within the cyber security architecture is an important first step in coming up with appropriate training objectives" [7, p. 223]. We aim to build upon this work by identifying the specific tasks within all roles that contribute to the cybersecurity of an organization so that future training can be designed to specifically target these tasks. This also allows for the opportunity to utilize job redesign and job crafting principles using the task information to enhance these tasks in situations where training may not be adequate.

Contextual AI Classification in Specialized Domains Large Language Models (LLMs) have advanced the classification of unstructured, complex datasets across different domains [8]. Several recent works have utilized LLMs to analyze the O*NET labor dataset [25, 4]. Although O*NET provides a structured representation of labor variables, mapping these variables into real-world applications is non-trivial. Recent studies in NLP have made significant progress in this area. Meisenbacher et al. [25] leveraged encoder-based architecture to extract the normalized O*NET features from large volumes of noisy job advertisement data. Parallel advancements have been adopted in the labor market analysis. Eloundou et al. [12] demonstrated efficacy of utilizing the LLMs to classify massive volumes of O*NET task descriptions to measure occupational exposure to generative AI.

A key challenge in this domain is the classification of O*NET tasks in isolation, where the absence of broader occupational context results in heightened semantic ambiguity. To overcome the limitation, we propose a two-stage classification pipeline detailed in Section 6.

4 Research Questions

The research questions we attempt to answer are as follows:

- **RQ1**: What is an appropriate definition of task-based cyber risk?
- **RQ2**: What task information is necessary to estimate task-based cyber risk?
- **RQ3**: How can machine learning or large language models be used to assist in categorizing tasks as containing task-based cyber risk?
  - **RQ3a**: What information does a model need to be provided to attain greater than 90% reliability for classifying task-based cyber risk when compared to human coders?
  - **RQ3b**: Does a dynamic allocation configuration that incorporates role information achieve high balanced accuracy while minimizing token expenditure and inference cost?

5 Manual Task Categorization Using Qualitative Analysis

5.1 Data Source for Tasks

The U.S. federal government through the Department of Labor created the Occupational Information Network (O*NET) database containing occupation profiles covering over 55,000 jobs across the economy [34]. Each occupation lists specific tasks. In total, 18,797 tasks are reported. Note that a single task can be used for multiple occupations. For example, the task to "design, or supervise the design of, systems, processes, or equipment for control, management, or remediation of water, air, or soil quality" is carried out by 20 occupations, such as Industrial Ecologists, Civil Engineers, Hydrologists and Conservation Scientists. We utilize the full task description as well as the list of roles related to the task.

5.2 Process for Manually Labeling Tasks

The procedure for determining whether job tasks exhibit cyber risk begins with constructing a definition that reflects the domain. We then put the definition to the test by using three independent human coders to label tasks over multiple rounds, refining the definition at each stage. The final result is a working definition of task-based cyber risk and a ground-truth dataset of labeled tasks that can be used to evaluate the automated approach described in the next section.

To begin, we agreed upon a definition of cyber risk based upon existing definitions provided by NIST SP 800-60 Vol. 1 Rev. 1 [24] and ISO 27005 [20].

Definition 1. Cyber risk *An effect of uncertainty on or within information and technology. Cybersecurity risks relate to the loss of confidentiality, integrity, or availability of information, data, or information (or control) systems and reflect the potential adverse impacts to organizational operations (i.e., mission, functions, image, or reputation) and assets, individuals, other organizations, and the Nation.*

With this definition of cyber risk we then drafted an initial definition of task-based cyber risk for the coders to apply and improve. Each coder is an experienced cybersecurity professional with varied backgrounds in cybersecurity and computer science. Following each round, the coders met and discussed any disagreements. In each round, a randomly sample of 150 tasks from the O*NET database were evaluated. Our intent through each round was twofold. First, we aimed to achieve acceptable coder agreement. Second, we sought to further refine our definition of task-based cyber risk for later use with a model for automated task classification. Five rounds were required to achieve acceptable coder agreement. The final definition is given below.

Definition 2. Task-based cyber risk *exists when one or more of the following holds: (i) performance of the task involves the access, use, modification, or deletion of sensitive or valuable data or systems; (ii) when tasks could fail due to the compromise of confidentiality, integrity, or availability of the associated data or systems; or (iii) when tasks could be directly targeted by attackers.*

Post-round discussions revealed that it is often difficult to assess whether a task contains cyber risk based solely on a yes or no scale. Table 1 illustrates the challenge with example tasks. In some cases, as in the top row of the table, the cyber risk is immediately apparent. In other cases, no cyber risk is involved, as in the second row of the table. But for many tasks, the distinction is not obvious. For example, "negotiate prices or terms of sales or service agreements" makes no mention of cybersecurity or IT infrastructure. Nonetheless, it does mention service agreements, which may be stored digitally, as well as negotiations, which could be conducted online. Moreover, the integrity and confidentiality of these negotiations is important and could be threatened by a cyber attack. Hence, the coders ultimately labeled the task as exhibiting cyber risk.

The final example in Table 1 illustrates a task that falls just short of constituting cyber risk. Again, the task involves communication and storage that

Table 1. O*NET Task Descriptions with Cyber Risk Determination

O*NET Task Description	Cyber Risk Determination
Maintain cyber defense software or hardware to support responses to cyber incidents.	Yes (5)
Administer traction to relieve neck or back pain, using intermittent or static traction equipment.	No (1)
Negotiate prices or terms of sales or service agreements.	Yes – Borderline (3)
Coordinate sales or other promotional strategies with merchandising, operations, or inventory control staff to ensure product catalogs are current, accurate, and organized for best findability against user intent.	No – Borderline (2)

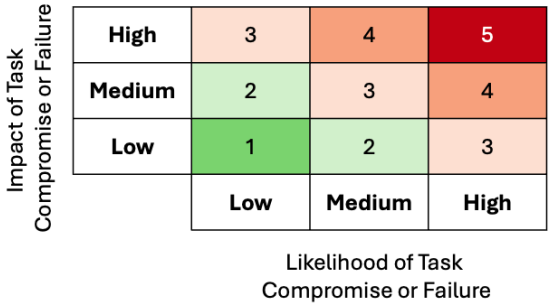

Fig. 1. Task-based Risk Matrix

could be digital. However, it is not especially plausible that a cybercriminal would profit from undermining the accuracy of product catalogs.

To systematically support such evaluations, we incorporated an additional risk classification to gauge the importance of any identified cyber risk. In rounds 3 and 4, the coders utilized the threat activities identified in the Cybersecurity Risk Foundation's Threat Taxonomy [11] to assess the severity of any prospective cyber risk. Ultimately, the method was determined to be too complex for the task of labeling tasks. Hence, in round 5 the coders adopted a simple approach based upon existing methodologies using a likelihood and impact rating for risk classification. Each task was rated on a matrix of the likelihood and impact rated low, medium, or high. These were then used to calculate total task risk. The matrix used is shown in Figure 1. While we did not provide strict guidance on what risk level would constitute task-based cyber risk, coders generally considered tasks that were scored either a one or two to not have cyber risk and tasks scored three or above to contain cyber risk.

Table 2. Inter-rater agreement statistics by round

Round #	Fleiss' Kappa	Gwet's AC1	% Agreement	# Tasks Rated	# Tasks Containing Cyber Risk
1	0.464	0.502	61.29%	300	136
2	0.585	0.861	84.40%	150	15
3	0.502	0.605	66.96%	150	38
4	0.458	0.745	62.67%	150	19
5	0.550	0.801	79.33%	150	30

5.3 Results for manual labeling

To assess inter-rater reliability of the coders after each round, we utilized multiple methods. First and most simply, we calculated the raw agreement percentage where all three coders agreed on a task classification. Next, we sought out measures of inter-rater reliability to understand if our agreement was due to random chance. Because we have more than two coders, Cohen's Kappa is not suitable [9]. Several other measures including Fleiss' Kappa and Krippendorf's Alpha can be used in certain cases when there are more than two coders; however, there are still issues with these measures in our case [14]. Throughout our initial discussions, all coders agreed that there are far more tasks that do not contain cyber risk than those that do, which can cause issues because, in these measures, the chance-correction baseline is inflated by the skewed category distribution [18]. To solve this problem, we turned to Gwet's AC1 which was created precisely to alleviate the issues with the previous measures [18, 37]. We set thresholds of 75% raw coder agreement and an AC1 of greater than 0.8. Those results can be seen in Table 2.

To assess the suitability of our definition for use with a large language model for automated classification, the coders met and discussed their ratings after each round. Any ambiguities identified in the definition during these discussions resulted in an updated definition. Following round three we began testing the definition with several models to gauge whether our definition was sufficiently tight. After five rounds we surpassed the established thresholds for both percent agreement and the AC1, and all coders found the definition to be suitable for our purposes.

6 Automated Task Categorization using LLMs

Having developed a refined definition of task-based cyber risk in the manual labeling, we incorporate the definition into an LLM-based approach. We first describe the process and then report on its accuracy compared to the ground truth data labeled by human coders.

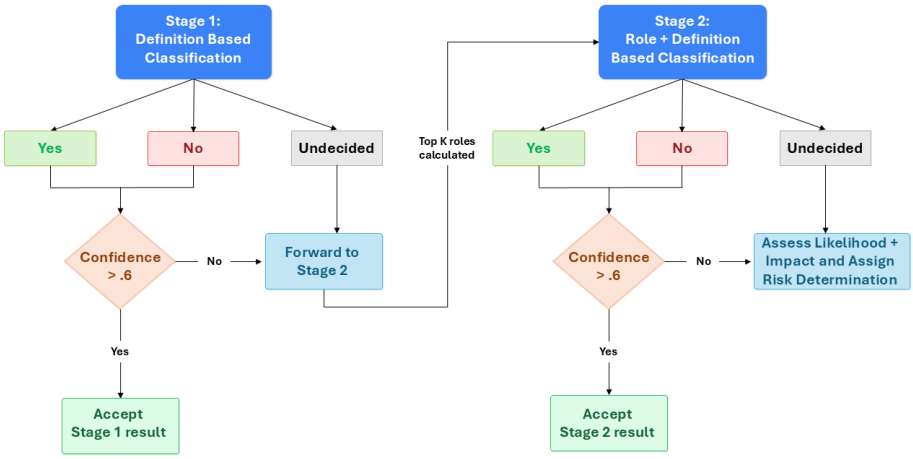

Fig. 2. Two Stage Automated Classifier Architecture

6.1 Process for automated labeling

In order to automate the categorization, a two-stage inference framework was developed as shown in Figure 2. In Stage 1, each task was mapped to the LLM through a structured zero-shot prompt containing the task-based cyber risk definition and task sentence. The LLM is instructed to classify whether the task satisfies the custom definition, producing a structured output consisting of definitive label (YES/NO/UNCERTAIN), a confidence metric [0,1], categorical scales for impact and likelihood (HIGH, MEDIUM, LOW, or N/A), a natural language explanation, and the specific definition clause matched. A confidence threshold θ_1 (set to 0.60 in the baseline configuration) is applied. Task sentences greater than θ_1 are accepted in Stage 1. Any classifications falling below θ_1, or those labeled with UNCERTAIN, proceed to Stage 2.

Task descriptions are often short and contextually ambiguous when classified in isolation. To overcome this limitation, a second inference stage leverages job role context in the form of associated role titles. Since the human coders decided to consider job roles in questionable cases, it makes sense for the LLM to do the same. These roles are ranked by semantic similarity using a 384 dimensional dense embeddings generated by *all-MiniLM-L6-v2* model. *All-MiniLM-L6-v2* is a sentence embedded model (Hugging Face) based on MiniLM distilled transformers [36] and trained using the Sentence-Transformers framework [30]. Cosine similarity between the task sentence and each role title embedding measures the ranking. To optimize computational efficiency the top K (K=3 in baseline configuration) roles are forwarded for further processing as each task could have many roles, and providing all roles would be computationally expensive and inefficient. These top K role titles with highest semantic similarity to the task description are injected into the prompt, along with similarity scores and rankings. This supplies the model with industry context that could resolve lexical ambiguities.

To resolve the instances that remain UNCERTAIN even after Stage 2 contextual escalation, the framework adopts a deterministic fallback strategy that once again mirrors what the human coders did. Here, the model utilizes the task-based risk scoring presented in Figure 1. When the calculated risk score for an UNCERTAIN task is rated as 3 or higher, it is labeled as having task-based cyber risk. If the score is 1 or 2, the task is deemed to not have cyber risk.

To determine the minimum model capacity required for this task and to evaluate the efficiency of capacity-differentiated routing, we tested the two-stage architecture across three distinct model configurations.

Configuration 1: Standalone Gemma3:4b This configuration is a strictly homogeneous and lightweight pipeline that tests how the additional semantic context can resolve low-confidence classifications without requiring heavy computational resources. In Stage 1, the sentences are processed through the 4 billion parameter *Gemma3:4b* [3] model. In Stage 2, task statements falling below the threshold confidence are given to the same *Gemma3:4b* with additional related roles. This establishes the baseline floor for token expenditure and classification accuracy.

Configuration 2: Standalone Llama3.1:8b To evaluate the framework with widely adopted open-source standards, we adopted a pipeline utilizing the 8 billion parameter *Llama3.1* [17] across both stages. This functions as a mid-tier evaluation point for both accuracy and inference cost, probing the performance characteristics of 8 billion parameter models under standardized architectural settings.

Configuration 3: Gemma3 Cascade This configuration tests the hypothesis that dynamic capacity allocation can achieve high balanced accuracy while minimizing token expenditure. Stage 1 acts as low latency filter, where minimal capacity *Gemma3:4b* model rapidly resolves the unambiguous tasks, there by reducing the inference costs. Stage 2 acts as high compute escalation protocol, tasks that are not resolved in prior stage are not only provided with additional occupational data but are also escalated with 12 billion parameter model *Gemma3:12b* [3]. This architecture is designed to route expensive, high token reasoning process solely to complex edge cases.

6.2 Results for Automated Labeling

All classification inferences were deployed locally using Ollama. The current study evaluates and compares the performance of three configurations. Gold labels were established from Round 5 of earlier human annotation, taking the majority vote among the YES/NO annotations as a reference standard. These gold labels are subsequently benchmarked against the two-stage classifier outputs to assess accuracy and reliability.

The dataset includes more tasks without cyber risk, reflecting its distribution in the overall population. Hence, we also utilize metrics that account for prevalence-sensitive reliability. To address the Kappa paradox associated with imbalanced datasets, both Cohen's Kappa and Gwet's AC1 were employed. In

Table 3. Comparison of LLMs performance

Metric	Config 1 **Gemma3:4b** (Standalone)	Config 2 **Llama3.1** (Standalone)	Config 3 **Gemma3** (Cascade)
Accuracy	0.76	0.84	0.79
YES Precision	0.45	0.65	0.50
YES Recall (Sensitivity)	0.53	0.43	0.67
NO Recall (specificity)	0.82	0.94	0.82
Balanced Accuracy	0.73	0.69	0.75
Cohen's Kappa	0.37	0.44	0.43
Gwet's AC1	0.70	0.81	0.74
95% CI (AC1)	0.61-0.79	0.74-0.89	0.65-0.82
False Negatives	9	16	9
False Positives	19	7	20
UNCERTAIN Rate	5.3%	0%	0.02%
Stage 2 Triggered	14	0	16
Stage 2 Resolved as YES	6	0	13
Stage 2 Remained UNCERTAIN	8	0	3
Accuracy after fallback strategy	0.78	0.84	0.80
Balanced Accuracy after fallback strategy	0.73	0.69	0.76

alignment with its application in calculating agreement among human annotators, AC1 was also used to measure the agreement and accuracy of automated label predictions. Given the absence of an UNCERTAIN class in the gold standard, model evaluations were conducted strictly against binary (YES/NO) ground truth labels. Therefore, we treated any UNCERTAIN model outputs as intentional abstentions.

Table 3 presents the comparison metrics of standalone *Gemma3:4b* (Configuration 1), standalone *Llama3.1* (Configuration 2) and cascading *Gemma3* (Configuration 3) performance. Standard evaluation metrics like Cohen's Kappa for all three configurations underestimated model agreement due to the imbalance of classes. The application of Gwet's AC1 indicates near-perfect agreement for both models. *Llama3.1* demonstrated the stronger agreement with gold labels, achieving an AC1 score of 0.81 with a 95% confidence interval ranging from 0.74-0.89.

Beyond baseline reliability, the models showed contrasting operational characteristics. Configuration 2 maximized accuracy and yes precision (0.65), whereas Configuration 3 demonstrated more responsive to the minority positive class, resulting greater recall (0.67) and balanced accuracy of 0.75. The number of false negatives and false positives remained almost same in both the configuration cases of standalone *Gemma3:4b* and *Gemma3* in cascading architecture. These findings suggest that Configuration 2 is the optimal model for minimizing the false positives, whereas Configuration 3 is better suited for maximizing the true positive detection and minimizing the false negatives for imbalanced datasets.

Furthermore, we observe an interesting trade-off between efficiency and accuracy tied to model uncertainty. Configuration 2 resolved all task statements at Stage 1, without requiring Stage 2. This demonstrates that *Llama3.1* consistently generated high-confidence predictions, eliminating the need for incorporating additional context about occupation and risk severity. This increases computational efficiency at the expense of dealing with uncertainty. In contrast, Configuration 1 triggered Stage 2 evaluation for 14 tasks. Of these, 6 were resolved as YES following the contextual augmentation. Configuration 3 routed 16 task statements to Stage 2, resolving 13 immediately.

Moreover, the fallback strategy utilizing the risk severity score helped further resolve the remaining UNCERTAIN tasks for Configurations 1 and 3. The resolution did not always translate to improved accuracy, however. For Configuration 1, accuracy went up slightly, but balanced accuracy remained unchanged. For configuration 1, the fallback strategy introduced 4 additional false negatives and 1 false positive. Conversely, for Configuration 3, the fallback strategy increased both overall accuracy and balanced accuracy. Notably, the risk-scoring technique has introduced only one false negative for Configuration 3.

When comparing Configuration 2 to 1 and 3, we observe that many of Configuration 3's false negatives correspond to initial UNCERTAIN ratings in Configurations 1 and 3. Hence, *Llama3.1*'s high-confidence negative predictions do not reliably signal absence of cyber risk, but instead reflect a bias towards conservative classification when statements lack explicit digital indicators. In contrast, *Gemma3*'s uncertainty helps trigger augmented contextual reasoning that the human coders also found valuable.

Taken together, these findings reinforce the decision to design the LLM in a manner that reflects the process emerging from multiple rounds of human-coding tasks. Bolstering definitions with additional context can help resolve uncertainty. The capacity-differentiated cascade successfully optimized the risk capture and inference cost. By leveraging the 4 billion parameter model for Stage 1 and utilizing the more computationally expensive 12 billion model for Stage 2, configuration 3 has achieved the highest balanced accuracy (0.75) and highest YES recall (0.67) of all tested pipelines. This experience illustrates that deploying massive models on every task is unnecessary. Dynamic compute escalation can sometimes resolve complex tasks while maintaining inference costs closer to lightweight baseline models.

7 Concluding Remarks

The results from this study are encouraging. Through our manual categorization we developed a definition of task-based cyber risk that is broad enough to capture tasks that may not be traditionally considered as risky by security practitioners (RQ1). While the definition was created with the intent to limit false negatives, we did not observe a significant increase in false positives among human coders. In the initial rounds of categorization coders were only provided the task description without additional information. This was found to be in-

sufficient to accurately judge whether a task contributed to cyber risk, so we also included the roles that were related to the task in question (RQ2). This addition allowed the coders to better understand the context in which the task is performed and therefore allowed for more accurate task categorization.

Our attempt to automate the categorization of tasks demonstrated the viability of using LLMs to assess whether tasks contain cyber risk. Though we achieved our initial goal of using an LLM to automate the assessment process (RQ3), the work is far from complete. We have yet to achieve our stated threshold (90%) of model accuracy, however we are optimistic that with additional prompt tuning we will be able to meet this goal (RQ3a). A comparison of the three model configurations indicated that utilizing a dynamic allocation configuration incorporating additional details did increase the balanced accuracy (RQ3b), by resolving the complex edge cases.

Our results indicate that the model that performs best will depend on the use case. If false negatives are not a concern, Llama3.1 may be sufficient. When false positives are more concerning the Gemma3 model should be considered. If inference costs are not a concern, both models may be used and compared to minimize the occurrence of both false negatives and false positives.

As articulated more fully in Section 2, we believe this methodology creates several exciting possibilities that should be explored in future research. One key area where this could be deployed is in the development of customized security awareness training. A full understanding of the specific cyber risks that are present within a role and the tasks that cause those risks would allow for custom training to be developed to target the task and related risks. Researchers should also explore whether this information can be utilized to help redesign tasks in a way that minimizes risk exposure. Finally, future work should explore whether task risk data can be used by industry sources like cyber insurance providers to create detailed risk profiles for organizations.

References

1. Adams, A., Sasse, M.A.: Users are not the enemy. Commun. ACM **42**(12), 40–46 (dec 1999). https://doi.org/10.1145/322796.322806, https://doi.org/10.1145/322796.322806
2. Ahmad, S., Alqaarni, S.: Job Analysis in Organizations: Transition From Traditional to Strategic. International Journal of Professional Business Review **8**(5), e01424 (May 2023). https://doi.org/10.26668/businessreview/2023.v8i5.1424
3. et al., G.T.: Gemma 3 technical report (2025), https://arxiv.org/abs/2503.19786
4. Athey, S., Brunborg, H., Du, T., Kanodia, A., Vafa, K.: Labor-llm: Language-based occupational representations with large language models. arXiv preprint arXiv:2406.17972 (2024)
5. Beautement, A., Sasse, M.A., Wonham, M.: The compliance budget: managing security behaviour in organisations. In: Proceedings of the 2008 new security paradigms workshop. pp. 47–58 (2008)
6. Bolger, C., Moore, T.: Expanding the scope: An empirical approach for identifying high-risk users. In: Workshop on the Economics of Information Security (WEIS) (2025)

7. Brummel, B.J., Hale, J., Mol, M.J.: Training cyber security personnel 1. Psychosocial dynamics of cyber security pp. 217–239 (2016)
8. Cherupalli, R., Grubbs, H., Chua, Y.T., Pei, W., Moore, T., Warner, G.: Contextual classification of cybercriminal posts using large language models: A comprehensive study on tech support scam marketplaces. In: 2025 APWG Symposium on Electronic Crime Research (eCrime). pp. 1–11. IEEE (2025)
9. Cohen, J.: A coefficient of agreement for nominal scales. Educational and psychological measurement **20**(1), 37–46 (1960)
10. CompTIA: What is information technology risk management? https://www.comptia.org/content/guides/what-is-information-technology-risk-management (2022), accessed: 2/12/2025
11. Cybersecurity Risk Foundation: Crf threat taxonomy, version 2026. Tech. rep., Cybersecurity Risk Foundation (2026), https://crfsecure.org/research/crf-threat-taxonomy/
12. Eloundou, T., Manning, S., Mishkin, P., Rock, D.: Gpts are gpts: Labor market impact potential of llms. Science **384**(6702), 1306–1308 (2024)
13. European Union Agency for Cybersecurity (ENISA): Enisa threat landscape 2025 (Oct 2025). https://doi.org/10.2824/2445233,
14. Feinstein, A.R., Cicchetti, D.V.: High agreement but low kappa: I. the problems of two paradoxes. Journal of clinical epidemiology **43**(6), 543–549 (1990)
15. Gordon, L.A., Loeb, M.P.: The economics of information security investment. ACM Trans. Inf. Syst. Secur. **5**(4), 438–457 (Nov 2002). https://doi.org/10.1145/581271.581274
16. Gordon, L.A., Loeb, M.P., Lucyshyn, W., Zhou, L.: Externalities and the magnitude of cyber security underinvestment by private sector firms: a modification of the gordon-loeb model. Journal of Information Security **6**(1), 24 (2015)
17. Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al.: The llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024)
18. Gwet, K.L.: Computing inter-rater reliability and its variance in the presence of high agreement. British Journal of Mathematical and Statistical Psychology **61**(1), 29–48 (2008)
19. Inglesant, P.G., Sasse, M.A.: The true cost of unusable password policies: password use in the wild. In: Proceedings of the sigchi conference on human factors in computing systems. pp. 383–392 (2010)
20. Information security, cybersecurity and privacy protection — guidance on managing information security risks. Standard ISO/IEC 27005:2022, International Organization for Standardization / International Electrotechnical Commission, Geneva, Switzerland (10 2022), https://www.iso.org/standard/80585.html
21. Knight, C., Parker, S.K.: How work redesign interventions affect performance: An evidence-based model from a systematic review. Human relations **74**(1), 69–104 (2021)
22. Kumaraguru, P., Cranshaw, J., Acquisti, A., Cranor, L., Hong, J., Blair, M.A., Pham, T.: School of phish: a real-world evaluation of anti-phishing training. In: Proceedings of the 5th Symposium on Usable Privacy and Security. pp. 1–12 (2009)
23. Marin, I.A., Burda, P., Zannone, N., Allodi, L.: The Influence of Human Factors on the Intention to Report Phishing Emails. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. pp. 1–18. ACM, Hamburg Germany (Apr 2023). https://doi.org/10.1145/3544548.3580985

24. McCallister, E., Grance, T., Scarfone, K.: Guide for mapping types of information and information systems to security categories. Tech. Rep. NIST Special Publication 800-60 Volume 1 Revision 1, National Institute of Standards and Technology, Gaithersburg, MD (August 2008). https://doi.org/10.6028/NIST.SP.800-60v1r1, https://doi.org/10.6028/NIST.SP.800-60v1r1, u.S. Department of Commerce
25. Meisenbacher, S., Nestorov, S., Norlander, P.: Extracting o* net features from the nlx corpus to build public use aggregate labor market data. arXiv e-prints pp. arXiv–2510 (2025)
26. MITRE Corporation: Mitre att&ck framework. https://attack.mitre.org/ (2024)
27. National Institute of Standards and Technology: Risk management framework (RMF). https://csrc.nist.gov/projects/risk-management/about-rmf (2020), 2/12/2025
28. Pfister, M., Apruzzese, G., Pekaric, I.: Department-specific security awareness campaigns: A cross-organizational study of hr and accounting. In: 2025 APWG Symposium on Electronic Crime Research (eCrime). pp. 1–17. IEEE (2025)
29. Ponemon Institute: Cost of a data breach report 2024. Tech. rep., IBM Security (2024), accessed: 2025-04-15
30. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP). pp. 3982–3992 (2019)
31. Rudolph, C.W., Katz, I.M., Lavigne, K.N., Zacher, H.: Job crafting: A meta-analysis of relationships with individual differences, job characteristics, and work outcomes. Journal of vocational behavior **102**, 112–138 (2017)
32. Schulze, L.J.H., Delclos, G.L., Pinglay, N.: Integrated Job Analysis: A Technique to Document Job Activities and to Identify Occupational Risk Factors and Modes of Remediation and Accommodation. International Journal of Occupational and Environmental Health **7**(3), 222–229 (Jul 2001). https://doi.org/10.1179/oeh.2001.7.3.222
33. Steves, M.P., Greene, K.K., Theofanos, M.F.: A Phish Scale: Rating Human Phishing Message Detection Difficulty. In: Proceedings 2019 Workshop on Usable Security. Internet Society, San Diego, CA (2019). https://doi.org/10.14722/usec.2019.23028
34. U.S. Department of Labor, Employment and Training Administration: About O*NET (2026), https://www.onetcenter.org/overview.html
35. Verizon: 2025 data breach investigations report. Tech. rep., Verizon (2025), accessed: 2025-04-15
36. Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., Zhou, M.: Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. Advances in neural information processing systems **33**, 5776–5788 (2020)
37. Wongpakaran, N., Wongpakaran, T., Wedding, D., Gwet, K.L.: A comparison of cohen's kappa and gwet's ac1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples. BMC medical research methodology **13**(1), 61 (2013)