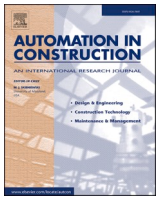
Automation in Construction 165 (2024) 105565

Contents lists available at ScienceDirect

Automation in Construction

journal homepage: www.elsevier.com/locate/autcon



Enhancing cyber risk identification in the construction industry using language models

Dongchi Yao [a,b,*], Borja García de Soto [a,b]

[a] S.M.A.R.T. Construction Research Group, Division of Engineering, New York University Abu Dhabi (NYUAD), Saadiyat Island, P.O. Box 129188, Abu Dhabi, United Arab Emirates
[b] Department of Civil and Urban Engineering, Tandon School of Engineering, New York University (NYU), New York 11201, United States

ARTICLE INFO

Keywords:
Cybersecurity
Risk identification
Deep learning
Language model
Construction industry

ABSTRACT

Modern construction projects are vulnerable to cyber-attacks due to insufficient attention to cybersecurity. Cyber risks in construction projects are not fully recognized, and the relevant literature is limited. To address this gap, the capabilities of a language model were leveraged to analyze extensive text, tailored to identify cyber risks. The model was trained using a curated corpus related to construction cybersecurity, enhanced by Supervised Fine-Tuning and Reinforcement Learning from Human Feedback techniques. The findings demonstrate advancements in the model's ability to understand cybersecurity and generate responses to cybersecurity questions. Using this model, a prioritized checklist of cyber risks across project phases was developed, establishing a new industry benchmark. This checklist can be utilized by various groups, including project managers and risk analysts. The model allows for updates with new data, ensuring the checklist remains current. The upgraded model holds significant promise for industry-wide applications, serving as an intelligent cybersecurity consultant.

1. Introduction

The construction industry has entered the digital era, known as Construction 4.0, which centers around cyber-physical systems and integrates technologies such as digital twins, drones, robotics, and virtual reality [1]. These technologies have improved the speed and quality of construction, operation, and maintenance of assets [2]. However, the construction industry significantly lags in cybersecurity awareness, making it a prime target for cyberattacks that could lead to project delays, financial losses, and other detrimental consequences [3]. This vulnerability has been highlighted by incidents such as Turner Construction falling victim to a spear-phishing scam [3], Marous Brothers Construction not receiving payment due to maliciously changed routing numbers [4], Bird Construction being breached by ransomware [5], and Hoffmann Construction reporting unauthorized access to employee information [6]. A key factor underlying these incidents, as noted in [7], is the generalized lack of awareness by industry stakeholders of the cyber risks that arise from threats (malicious actions aimed at exploitation) and vulnerabilities (specific weak points susceptible to exploitation) inherent in construction projects [8]. This lack of recognition results in the absence of a benchmark checklist and, consequently, a lack of comprehensive preventive measures, which are crucial for protecting sensitive data and maintaining operational integrity.

Therefore, it is imperative to clearly identify cyber risks, including threats and vulnerabilities. Doing so can raise the awareness of project managers in construction companies and facilitate the implementation of preventive and proactive measures against cyber attacks. To ensure broad relevance and applicability across construction projects, these cyber risks can be categorized according to the different project phases: initiation, design, construction & procurement, commissioning, operation & maintenance, renovation, and end of life. This categorization, as recognized in the literature [7], reflects the universal progression of construction projects, regardless of their size, location, or delivery method. Aligning risk identification with these phases enables the construction industry to recognize and tackle the widespread challenges of cybersecurity more effectively.

However, existing studies on cyber risk identification in the construction industry are limited and do not provide a comprehensive, industry-specific list of cyber risks that could serve as a credible benchmark. Furthermore, the methods used have several drawbacks: they rely heavily on manual implementation, which can introduce bias and oversights due to the varying levels of expertise among the

* Corresponding author at: Department of Civil and Urban Engineering, Tandon School of Engineering, New York University (NYU), New York 11201, United States.
E-mail addresses: dongchi.yao@nyu.edu (D. Yao), garcia.de.soto@nyu.edu (B. García de Soto).

https://doi.org/10.1016/j.autcon.2024.105565
Received 16 October 2023; Received in revised form 13 June 2024; Accepted 13 June 2024
Available online 24 June 2024

implementers.

Recently, the surge in natural language processing (NLP) techniques has led to an increased use of text analysis for risk identification and management across various industries, as shown by [9–11]. Considering the abundance of textual sources related to cybersecurity and construction, such as news articles, academic papers, and published reports [12]—both existing and forthcoming—these can be meticulously analyzed to identify emerging cyber risks in construction. Inspired by Large Language Models (LLMs) like Baidu's Ernie Bot [13], OpenAI's GPT-4 [14], and Google's Gemini [15], LLMs trained with Supervised Fine-Tuning (SFT) and Reinforcement Learning from Human Feedback (RLHF) techniques can synthesize and analyze these abundant text sources, which allows to identify cyber risks comprehensively. In the future, the developed model can continue to receive periodic training with new text data. This will ensure it remains aligned with the evolving cybersecurity landscape in construction and facilitates regular updates to the identified cyber risks, enabling dynamic cyber risk identification. Continuous training with new data is more efficient than traditional methods, which require starting from scratch with human intervention, thereby enhancing efficiency [16].

The goal of this study is to comprehensively identify cyber risks across project phases, achieved by innovatively developing a language model dedicated to construction cybersecurity. This study is structured around three objectives: (1) To train a base model that encodes knowledge of both cybersecurity and construction, demonstrating potential scalability by training efficiently on a single GPU. (2) To improve the model's ability to understand and generate answers about cybersecurity content and questions by implementing SFT and RLHF techniques. (3) To compile a comprehensive cyber risk checklist categorized by project phases, which will serve as a new benchmark that can be referred to for the formulation of preventive measures. It should be reiterated that the identified cyber risks are applicable to diverse projects rather than specific ones, and the detailed risk mitigation strategies for each risk are beyond the scope of this study.

The rest of the paper is structured as follows: Section 2 presents a literature review of language models, as well as existing studies on cyber risk identification in the industry. Section 3 details the methods for data creation, model development, cyber risk identification, and evaluations. Section 4 presents the implementations and results. Section 5 offers a thorough discussion of the identified cyber risks and the developed language model. Section 6 summarizes the main findings, clarifies the contributions, and discusses the limitations and areas of our future works.

2. Related works

In general, existing construction-related studies in cybersecurity do not support dynamic risk identification in evolving scenarios, as they require starting from scratch for each new assessment, thus lacking adaptability and flexibility. Additionally, current approaches are time-inefficient, necessitating lengthy discussions and checks among implementers. Given the construction sector's growing dependency on computational tools and the rapidly evolving cybersecurity landscape, it becomes imperative to shift to an automated approach, which needs to offer greater comprehensiveness, dynamism and efficiency.

2.1. Cybersecurity challenges in construction

Cybersecurity in construction differs from general practices due to the sector's dynamic nature, characterized by frequent changes in team composition, including subcontractors, architects, engineers, and site managers, leading to varied levels of cybersecurity expertise. Construction projects rely on robust communication across the supply chain, necessitating the exchange of sensitive information such as design specifications, schedules, and financial data via digital platforms. This increases vulnerability to data breaches and cyberattacks, highlighting the need for specialized cybersecurity measures in the construction industry. Stakeholders in the construction industry often share digital data, including Building Information Modeling (BIM) files and project management documents. BIM, which features detailed 3D models of buildings and infrastructure, is particularly attractive to cybercriminals seeking to exploit intellectual property or disrupt projects. Furthermore, the overlapping nature of construction projects, with multiple sites or phases managed concurrently by a single company or various teams, creates a complex network of digital interactions. This complexity can challenge the consistent implementation of cybersecurity measures across operations, underscoring the need for a cybersecurity strategy specifically tailored to the construction sector's unique challenges [17–19].

2.2. Large language models

As indicated in Section 1, LLMs have the capability to analyze abundant text related to construction cybersecurity, so that it has great potential for identifying cyber risks through text analysis. Language modeling is typically represented as the probability of a sequence of words $(w_1, w_2, \ldots, w_n)$ to represent the joint probability of a sentence. Utilizing the chain rule of probability, this concept is broken down into a series of conditional probabilities, as demonstrated in Eq. (1) [20].

$$P(w_1, w_2, \ldots, w_n) = P(w_1) \times P(w_2|w_1) \times P(w_3|w_1, w_2) \times \ldots \times P(w_n|w_1, w_2, \ldots, w_{n-1}) \quad (1)$$

Language models aim to maximize the probability of sentences that appear in the ground truth corpus dataset. This enables them to generate sentences that fall within the same distribution as the dataset, which has been performed well by modern LLMs. The growth of LLMs has significantly influenced the field of NLP, with applications extending across various industries. This development can be traced back to the introduction of the transformer architecture in 2017 by Vaswani et al. [21]. The transformer's self-attention mechanism and parallel computation enabled the training of more semantically rich word embeddings. In 2018, OpenAI introduced GPT [22], a transformer-based architecture trained via unsupervised pre-training and fine-tuning for specific NLP tasks. Google's BERT [23], also introduced in 2018, utilized transformer architecture and improved performance on many NLP tasks. In 2019, OpenAI released the larger GPT-2 [24] with remarkable language generation capabilities, and in 2020, they launched GPT-3 [25], notable for its zero-shot/few-shot learning ability.

Since the introduction of GPT-3, LLMs with billions of parameters have been developed, showing strong performance across many tasks, such as language generation, question answering, and machine translation. LLMs are significantly expanding AI's impact in various domains, including healthcare (e.g., medical image analysis, drug discovery), gaming (e.g., game development, chatbots), finance (e.g., fraud detection, risk management), robotics (e.g., natural language interaction, sensor data analysis), and enterprise software development (e.g., code completion, testing). In 2023, OpenAI introduced its latest model, GPT-4 [14], a large multimodal model capable of accepting text and image inputs and generating text outputs, exhibiting human-level performance

on various benchmarks, and offering improved reliability and creativity over its GPT-series predecessor. In addition to GPT-4, numerous other LLMs have been developed, including ChatGPT [26], PaLM [27], and LaMDA [28].

The SFT and RLHF techniques represent an important milestone for LLMs, as they enhance their capabilities by incorporating human preferences into the fine-tuning process through the use of reinforcement learning. RLHF incorporates human evaluations to align models with nuanced human values. An initial language model is pre-trained on textual data and fine-tuned using a reward model generated from ranked human feedback. The optimization process employs algorithms like Proximal Policy Optimization (PPO) [29] or the REINFORCE algorithm [30] to maximize alignment with human preferences. SFT and RLHF techniques have been successfully applied in models like GPT-4 (OpenAI) [14], Gopher (DeepMind) [31], and Anthropic [32], resulting in improved performance in natural language tasks. Through iterative updates, the techniques enable more human-like and reliable language models for diverse applications.

2.3. Cyber risk identification in construction

Several cybersecurity standards and tools, such as the National Institute of Standards and Technology (NIST) [33], General Data Protection Regulation (GDPR) [34], ISO/IEC 27000 [35] and Center for Internet Security (CIS) Controls [36] offer frameworks for managing cybersecurity risks. However, these are not specifically designed for the construction industry. Research tailored to construction cybersecurity is limited and scattered in topics. In 2023, a scoping review [19] identified 45 documents related to construction cybersecurity, including 25 peer-reviewed articles, 12 conference proceedings, and various other publications. These studies can be categorized into general discussions (24 documents), review papers (2 documents), and specific solutions (19 documents). General discussions feature works by Bello and Maurushat [37], El-Sayegh et al. [38], García de Soto et al. [1], Mantha and García de Soto [17], Yao and García de Soto [39], among others. Review papers include works by Pärn and Edwards [18], Sonkor and García de Soto [40], and Goh et al. [41], to name the most relevant ones. Specific methods or solutions encompass blockchain technology [18,42], machine learning or deep learning algorithms [18,43,44] threat modeling [7,45], framework proposals [8], or using the common vulnerability scoring system (CVSS) [46–48].

Additionally, very few studies related to cyber risk identification within the construction industry were found. For example, Shemov et al. [42] examined blockchain applications in construction supply chains, studying its impact on a case study and introducing a threat analysis model for potential attacks and countermeasures. Shibly and García de Soto [45] developed a construction-focused threat modeling approach, applying it to an offsite 3D concrete printing system to reveal vulnerabilities and potential countermeasures. Mantha et al. [7] presented a preliminary cybersecurity threat model for the AEC (Architecture, Engineering, and Construction) industry, illustrating its feasibility with an example from the commissioning phase of a building. However, these methods are manually established, which is time-consuming and may incur bias. Furthermore, they are not updatable over time, failing to reflect the evolving nature of cybersecurity threats and vulnerabilities. This underscores the need for a more automated, objective, and dynamic solution for cyber risk identification.

In summary, given the insufficient recognition of cyber risks in various construction projects, the limited literature where most methods require significant human involvement, the volume of construction cybersecurity-related text data both existent and expected, and the inspiration from LLM applications in other industries such as finance [49], this study aims to develop a language model for identifying cyber risks across project phases.

3. Methods

Our language model development is based on the framework described in [50], where a language model named InstructGPT was trained by implementing SFT and RLHF techniques. By incorporating human feedback, these techniques enhance the model's ability to generate text that aligns more closely with human expectations. Building on this, our study adapts the SFT and RLHF techniques with the goal of training a language model that can understand cybersecurity content

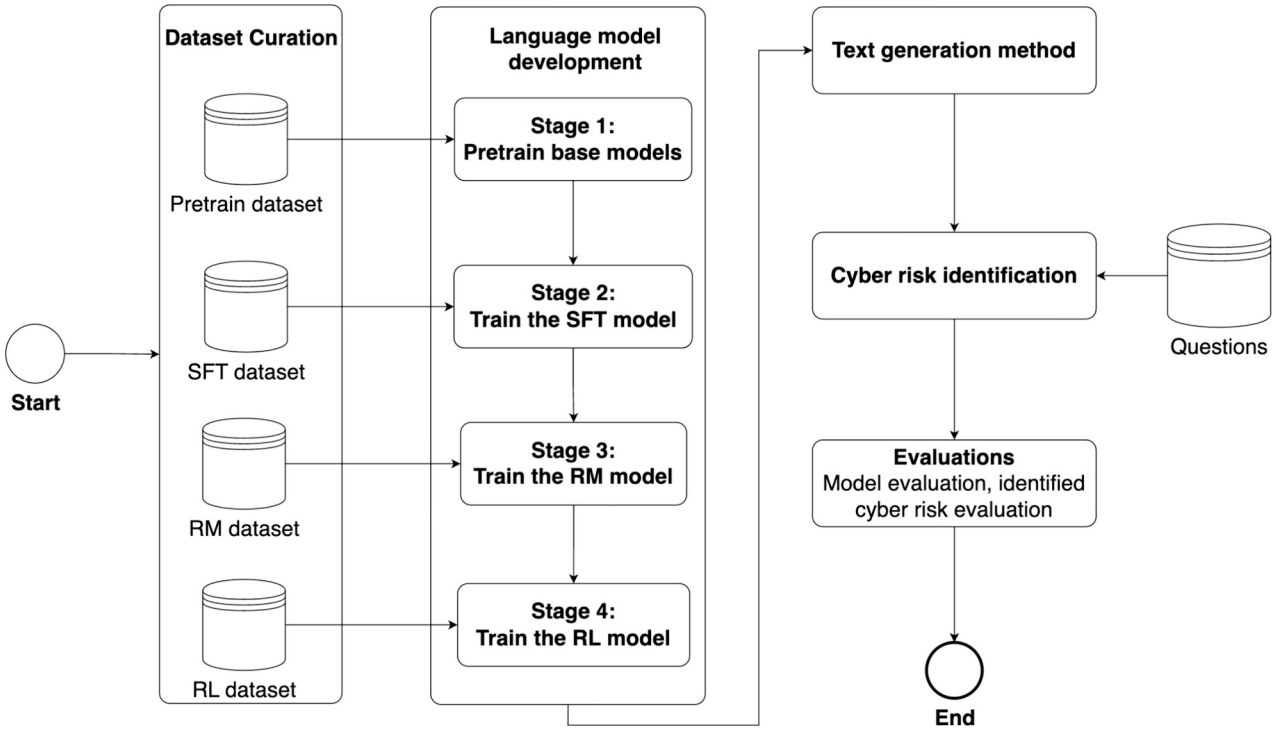


Fig. 1. Flowchart with main elements of the model development process.

and generate answers to cyber risk identification queries.

As illustrated in Fig. 1, the model development process encompasses four stages.

Stage 1: Collect a large corpus and pretrain a base model. We initially identified a substantial text corpus (pretrain dataset) on cybersecurity and construction. Then, we pretrained a language model on this corpus to encode the domain knowledge. This pretrained model serves as the base model.

Stage 2: Compile a question-answer dataset and train a supervised model (SFT model). We have developed a dataset of question-answer pairs related to construction cybersecurity (SFT dataset) to fine-tune the base model, which allows the model to capture diverse question formats and styles and results in our SFT model.

Stage 3: Compile a scoring dataset and train a reward model (RM model). We compiled a dataset of question-answer pairs, where each label represents a score indicating the degree of human preference for the answer to the question (RM dataset). A reward model, built on the SFT model, was trained to emulate the human scoring. This reward model played an important role in the RLHF technique implementation in the next stage.

Stage 4: Compile a question dataset and train a reinforcement learning model (RL model). We compiled a new dataset (RL dataset) and used the RM model to score answers from the RL model. This score was then used to fine-tune the RL model with a specially designed loss, enhancing its generalization to unseen prompts. This resulted in the final RL model.

After training the RL model, it is employed to generate text for answering questions related to cyber risk identification. From the answers, we compile a list of cyber risks applicable to various phases of a construction project. The evaluation includes both the assessment of the model's performance progress and the scrutiny of identified cyber risks. Specifically, the model evaluation concentrates on the model's advancement through different stages in understanding construction cybersecurity content and generating responses to cybersecurity questions. The evaluation of the identified cyber risks involves diverse sources, such as comparisons with existing literature, expert evaluations, and evaluations by the state-of-the-art GPT-4 model.

3.1. The dataset for each stage

Four datasets were utilized, each related to either the construction industry or cybersecurity. The pretraining dataset is derived from our previous work [12], while the other three datasets were carefully curated by the authors over a four-month period. To ensure alignment with each model's objectives, all prompts underwent rigorous scrutiny.

Table 1
Details of the pretrain dataset.

Text source	Descriptions	Sentence count	Total count
News articles, blogs	Cover incidents, causes, processes, and outcomes; offer timely insights and diverse perspectives	6 K	66 K
LexisNexis database	Provides legal, business, news, and records; regularly updated for current information	49 K	
Academic publications	Deliver professional knowledge, emphasizing relevance and rigor across various research fields	2 K	
Books (chapters)	Comprehensive domain knowledge for in-depth understanding through detailed analyses	6 K	
Specifications/ Standards	Detail cybersecurity legal/industry requirements, ensuring compliance and best practices	2 K	
Company reports	Reflect industry insights, strategies, and outcomes from a corporate perspective	1 K	

3.1.1. Pretrain dataset for Stage 1

The pretrain data, from our previous work [12] comprises 66 K sentences from a large text corpus spanning six cybersecurity-related sources in the construction industry, as detailed in Table 1. The corpus underwent semantic screening, using a BERT classifier to assess each sentence's relevance to construction cybersecurity by assigning a probability to its semantic suitability for the dataset. In this study, we excluded sentences assigned a weight not greater than 0.75 by the BERT classifier. This decision was based on our observation that probabilities below this threshold often approached or fell below 0.5, indicating the BERT classifier's low confidence in the semantic relevance of these sentences. This filtering left us with 61,841 sentences, constituting 93.7% of the original dataset. Recognizing the higher quality of academic writing, sentences from academic publications and book chapters were weighted at 1.5, making up 32.9% of our corpus. Therefore, each sentence thus received two weights: one for semantic quality from the BERT classifier and another for academic value. These weights were integrated into the loss function of the base model shown in Eqs. (2) and (3).

3.1.2. SFT dataset for Stage 2

For the SFT stage, we curated 326 questions related to the construction or cybersecurity domain, each with a complete answer. To enable the model to adapt to different linguistic styles, each question was rephrased into four alternative sentences, capturing both interrogative and imperative forms. The interrogative form typically ends with a question mark and is used for seeking information, while the imperative form usually ends with a period and is used for giving commands or requests. This allows the model to recognize the same query irrespective of its linguistic presentation and generate the same answer, enhancing its adaptability. Therefore, the dataset comprises $326 \times 5 = 1630$ question-answer pairs. 5 example question-answer pairs related to the cybersecurity domain are shown in Table 2, and the complete dataset can be found on our GitHub page [51].

3.1.3. RM dataset for Stage 3

The RM dataset initially consisted of 150 questions. For each question, we used the SFT model to generate four answers, resulting in a total of 600 question-answer pairs. The format of each pair is: "Question: {}; Answer: {}". Each pair was scored from 0 to 100 based on following criteria: Incomplete Answer (0−20) for answers missing essential information; Mention of Relevant Terms (20–40) for answers with related terms but lacking depth; Partially Related Answer (40–60) for answers somewhat addressing the question but missing key details; Logical Answer (60–80) for answers that are logical but not exhaustive; and Comprehensive and Logical Answer (80–100) for fully detailed and logically sound answers. To reduce bias, we held two scoring sessions spaced two days apart, presenting question-answer pairs in random order each time. The final score was the average of these two sessions. The RM dataset is available on our GitHub page [52].

Table 2
SFT dataset examples.

Question	Answer
How are cyber threats identified? How are cyber threats commonly detected? Can you explain how cyber threats are identified? Describe the process of identifying cyber threats. Explain the methods for identifying cyber threats.	Cyber threats may be identified through various means, including threat intelligence, security information and event management (SIEM), and user behavior analytics.

3.1.4. RL dataset for Stage 4

We curated 353 questions (without answers) related to the construction industry or cybersecurity, which are different from those in the SFT dataset. These questions, created also in either interrogative or imperative form, were used for training the RL model and can be found on our GitHub page [53].

3.2. The model for each stage

This section presents the pretraining, SFT, and RLHF techniques for model training. Details of the model training and selections will be provided in Section 4.1. Model progress evaluation methods and results are provided in Sections 3.5.1 and 4.2, respectively.

3.2.1. Base model for Stage 1

Pretraining a base model with a domain-specific corpus is crucial as it serves as a strong starting point for encoding cybersecurity knowledge and insights. In [50], the GPT-3 model, which has 175 billion parameters, was chosen as the base model. However, due to limited computing resources and the relatively smaller size of our dataset, we experimented with three types of smaller models.

- GPT-2: A generative language model by OpenAI, GPT-2 [24] is notable for human-like text generation, widely applied for task automation and insights generation. It is a precursor to GPT-4 [14], utilizing an autoregressive mechanism for word prediction based on prior context, and is pretrained on a comprehensive dataset.
- BERT-LM: Google's BERT [23] is a powerful transformer model acclaimed for its deep semantic understanding through bidirectional context analysis. We reset the configuration to make it only see the prior context and not what follows, and added a linear layer for autoregressive language modeling.
- T5-LM: Google's T5 [54], an encoder-decoder model, excels at translation and summarization via a uniform text-to-text approach. It features a BERT-like encoder and a GPT-2-like decoder. Our focus is on the decoder for language modeling, setting the encoder's input as "Language modeling task" without updating its weights.

The typical autoregressive training objective for language models uses cross-entropy loss [20,55]. As stated in Section 3.1.1, considering the weight indicating each sentence's semantic relevance and academic value, we integrate the two weights into a combined weight (Eq. (2)), which is then incorporated into the pretraining loss function in Eq. (3) to scale the loss of the corresponding sentence.

$$w_i = \frac{w_{i,w_1} \bullet w_{i,w_2} \bullet 1_{academic(i)}}{\min\limits_{1 \le i \le N} \left(w_{i,w_1} \bullet w_{i,w_2} \bullet 1_{academic(i)} \right)} \tag{2}$$

$$L_{weighted} = -\frac{1}{batch\_size} \sum_{i=1}^{batch\_size} w_i \sum_{k=1}^{n_i} \sum_{c=1}^{|V|} y_{i,k,c} log(p_{i,k,c}) \tag{3}$$

where $|V|$ is the vocab size for each model; w_{i,w_1} is the original weight assigned to the i-sentence; w_{i,w_2} is the weight for sentences from academic sources, initially set to 1; w_i is the relative combined weight for the i-sentence; $1_{academic(i)}$ is a function that returns 1.5 if the i-sentence is from an academic source, and 1 otherwise; n_i is the number of words in the i-th sentence; $y_{i,k,c}$ is 1 if the ground-truth label for the k-th word in the i-th sentence is the c-th class of word, otherwise 0. The ground-truth label is the next word of the k-th word; $p_{i,k,c}$ is the predicted probability that the next word of the k-th word in the i-th sentence is the c-th class of word.

3.2.2. SFT model for Stage 2

Our ultimate goal is triggering the model to generate answers that highlight possible cyber risks, given a question or prompt. Although the base model has encoded construction cybersecurity knowledge, the model should be aligned with our final downstream task so that it can understand the specific requirements of the task and produce more relevant and accurate outputs [50]. To this end, we implemented SFT technique to further train the model, an approach proved effective by various LLM studies [50,56,57]. This process involves fine-tuning our base model using our customized question-answering pair dataset, which is the SFT dataset in our study. It includes question-answer pairs in various formats, styled to resemble diverse styles of questions asked by different human, that the model is required to recognize. We concatenated each pair of questions and answers with a "\n" token in between to make it a coherent string available for the model's autoregressive training [50].

3.2.3. RM model for Stage 3

To increase the model's generalization ability of answering diverse questions beyond those in the SFT dataset, the RLHF technique [30] can be implemented. This involves providing feedback to the generated answers and optimizing the model against the feedback. In InstructGPT [50], the authors trained a reward model as the feedback provider, which is built on its SFT model and trained to predict a scalar value as the reward feedback. They designed a comparison mechanism between two different answers to the same question, which is incorporated into the loss function. Similarly, we added a regression head to our SFT model to create the reward model. However, due to the smaller size of our model and dataset, we adopted a more straightforward approach for reward model training. The model is trained to predict a score on a continuous scale from 0 to 100, with the objective of minimizing the difference between the predicted and the ground truth score labeled by humans.

3.2.4. RL model for Stage 4

Following [50], we employed the RLHF technique to further fine-tune the SFT model. This step aims to enhance the model's ability of generalization to unseen questions. Fig. 2 illustrates the RL fine-tuning process. Each question from the RL dataset is processed by the RL model to generate an answer, and the resulting question-answer pair is then evaluated by the RL model, the reward model, and the SFT model. Additionally, some data from the pretrain dataset are also processed by the RL model. These steps result in a loss function with three terms as shown in Eq. (4), adapted from [50].

$$l_{RL} = -w_1 \bullet \frac{1}{T} \sum_{i=1}^{T} \left(log(a_i) \bullet \frac{Score(x,y)}{100} \right) + w_2 \bullet KL(RL(y|x), SFT(y|x)) + w_3 \bullet E_{x' \sim D_{pretrain}} Pretain(x') \tag{4}$$

(1) **Score term.** This term incorporates the reward output by the RM model, evaluating the quality of the generated answer. This reward is normalized to fall between 0 and 1 and then multiplied by the log probability (a_i) associated with each token (word) in the generated answer. The resulting product is averaged over the

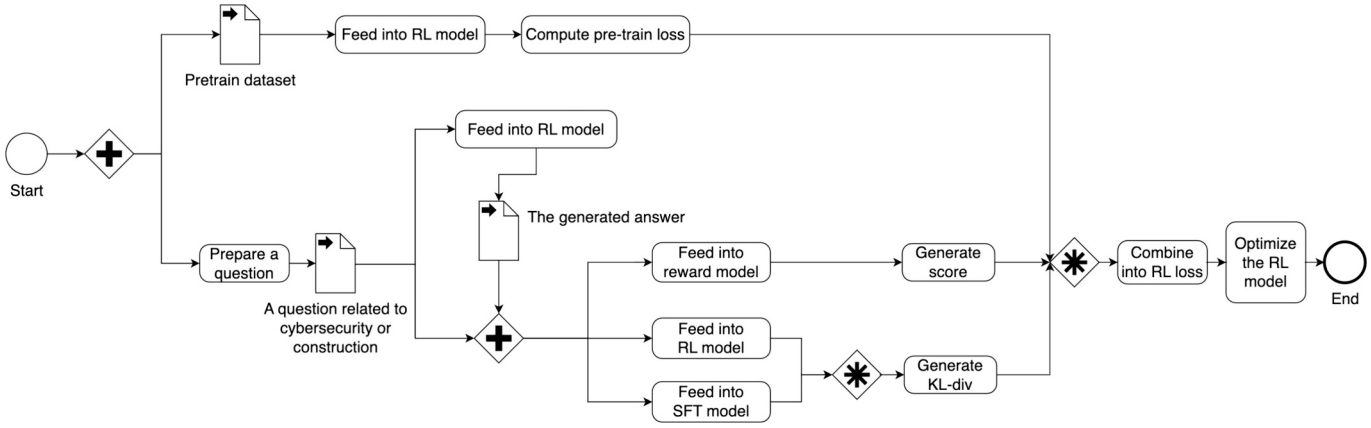


Fig. 2. Overview of RL fine-tuning.

sentence length, T. This term is consistent with REINFORCE algorithm, a standard policy gradient optimization method [30].

(2) **KL term.** The per-token Kullback-Leibler (KL) divergence penalty, shown in Eq. (5) [58], measures the discrepancy between the probabilities of sentences generated by the RL model and the SFT model. This penalty ensures that the answers generated by the RL model remain consistent with the knowledge acquired during the SFT stage, promoting more coherent answers. Additionally, it mitigates the risk of the RL model over-optimizing to the reward model, which could lead to the generation of nonsensical content [50].

$$KL(P_1, P_2) = \frac{1}{N}\sum_{i=1}^{N}\left(\sum_{j\in V} P_{1i}(j) log\left(\frac{P_{1i}(j)}{P_{2i}(j)}\right)\right) \tag{5}$$

where $P_{1i}(j)$ is the probability of choosing the j-th token at the i-th

Table 3
Pseudo-code for text generation.

	Algorithm: Autoregressive Text Generation
1:	BEGIN
2:	Set MAX_LENGTH and MAX_RETRY
3:	LOAD MODEL AND TOKENIZER
4:	Load the GPT-2 model architecture using "transformers" library from Hugging Face's repository
5:	Load the GPT-2 tokenizer using the same library
6:	Load specified weights to the model for initialization
7:	END LOAD MODEL AND TOKENIZER
8:	ENCODE SEQUENCE
9:	Initialize the question
10:	Tokenize the question using GPT-2 tokenizer
11:	Initialize encoded_sequence
12:	For each token in the tokenized question
13:	Convert the token to its numerical index in the GPT-2 vocabulary
14:	Append the index to encoded_sequence
15:	End For
16:	Return encoded_sequence
17:	END ENCODE SEQUENCE
18:	GENERATE TOKENS
19:	Initialize token_sequence with the encoded sequence
20:	Set error_count to 0
21:	While length of token_sequence < MAX_LENGTH
22:	Try
23:	Predict the next token using beam search
24:	Append the predicted token to token_sequence
25:	Catch SpecificError
26:	Increment error_count
27:	If error_count >= MAX_RETRY, then return Error
28:	Else
29:	Handle the error or resample to attempt a new prediction
30:	End If
31:	End Try
32:	End While
33:	Return token_sentence
34:	END GENERATE TOKENS
35:	REVIEW DECODED SENTENCE
36:	Decode the token_sequence using the tokenizer
37:	If the decoded sequence is satisfactory, then return the decoded sentence as the answer
38:	Else
39:	Call GENERATE TOKENS again
40:	End If
41:	END REVIEW DECODED SENTENCE
42:	END

position in the generated answer by RL model, $P_{2i}(j)$ is the probability of choosing the j-th token at the i-th position in the generated answer by the SFT model, N is the number of tokens in the generated answer, and V is the set of all possible tokens in the vocabulary.

(3) **Pretrain term.** The pretrain term also helps constrain the RL model to avoid over-optimizing, ensuring the model retains its generative capability aligned with the distribution of the pre-trained dataset. The inputs are minibatches from the pretrained dataset, and the mean cross-entropy loss within a minibatch is computed. Due to computing resource limitations, we only extract 20% of the sentences from the pretrained dataset with the highest combined weights (ensuring the model can still capture the distribution of the most weighted sentences) and divide them into a number of batches equal to the length of the dataset for the RL model.

3.3. Autoregressive text generation

Through this study, the standard algorithm for text generation is employed, which involves generating each word autoregressively [14,20,24,28,54]. The maximum length of tokens is to be 40 including the question part, which is deemed reasonable for a long sentence for the GPT-2 tokenizer [24]. Moreover, beam search [59] was adopted during the generation process to ensure the generated sentences have high joint probability. The pseudo code is shown in Table 3.

3.4. Cyber risk identification

The final RL model was tasked with identifying cyber risks across construction project phases. This was achieved by inputting crafted questions into the model, which then generated answers. To align the questions with the styles used in our SFT dataset, various phrasing structures were crafted, as shown in Table 4. To ensure the model could recognize different expressions of cyber risks and phases, we varied the keywords used within these questions. According to [8], risk identification includes threat and vulnerability identification, so the first set of keywords, {key1}, encompasses "cyber risks", "threats", and "vulnerabilities". The second set, {key2}, encompasses terms corresponding to different project phases, as shown in Table 5. This classification of phases aligns with the one in previous work on threat modeling [7]. The generated answers were subsequently reviewed and transformed into a checklist of identified cyber risks, ensuring the contextual accuracy and quality.

For each phase, the likelihood of each risk was derived from the RL model. We first formatted an identified risk, r, into a prompt: '{key1} in the {key2} phase includes {r}'. The prompts together make up the prompt set S_r for the phase. Then, the average probability over the number of tokens and the number of prompts in the set is computed, presented as Eq. (6). The probability of all risks in the phase was normalized to a range from 1 to 5 using min-max normalization [55] for later comparisons with other assessors [60].

$$p(r) = \frac{1}{|S_r| \bullet T} \sum_{s \in S_r} \sum_{t=1}^{T} p\left(x_{s,t} \middle| x_{s,1:t-1}\right) \tag{6}$$

Table 4
Structuring of cyber risk identification questions.

No.	Phrasing
1	What are the {key1} in the {key2} phase?
2	Can you identify {key1} in the {key2} phase?
3	{key1} in the {key2} phase include
4	Identify {key1} in the {key2} phase.
5	What types of {key1} should be considered during the {key2} phase?

Table 5
Different expressions for project phases (Key2).

Phase	Different terms
Initiation	pre-planning, concept, feasibility study, conception, early project definition, inception
Design	design development, schematic design, detailed design, planning
Construction &Procurement	construction, procurement, build, execution, implementation
Commissioning	handover, startup, completion, closeout
Operation & Maintenance	operation, maintenance, operations phase, facility management, service phase
Renovation & End of Life	renovation, end of life, demolition, decommissioning, retrofitting, rehabilitation, disposal

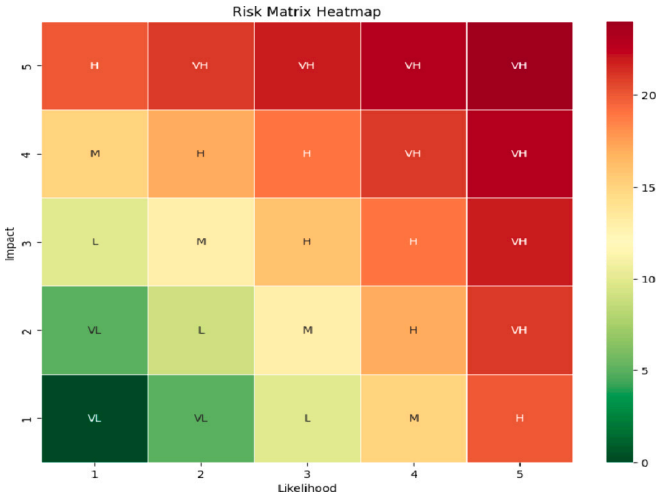


Fig. 3. The 5 × 5 risk matrix.

where T is the token count of prompt s, $x_{s,t}$ is the t-th token of s, $x_{s,1:t-1}$ are all tokens before the t-th token.

Detailed in Section 3.5.2, two industry experts and the GPT-4 [14] served as additional assessors to give likelihood assessment on a 5-point Likert scale [60–62], aiming to compare with our RL model for evaluation. These additional assessors were also requested to assess other attributes including relevance, impact, and consequence. Finally, for each risk, the likelihood and impact levels were averaged across all assessors. The risk value is calculated as the product of the averaged likelihood and impact levels [62,63]. This value was then categorized into five categories according to the risk matrix adapted from the literature [62–64] as shown in Fig. 3: Very Low (VL, 0–5), Low (L, 6–10), Medium (M, 11–15), High (H, 16–20), and Very High (VH, 21–25).

3.5. Evaluations

This study employs a two-layered approach for evaluation: (1) evaluating the model's progress, demonstrating its suitability for the task of identifying cyber risks, and (2) evaluating the compiled list of cyber risks, validating its applicability. The methods are multifaceted and incorporate both quantitative and qualitative approaches.

3.5.1. Evaluating model progress

As language models are fundamentally focused on understanding and generating text [16], we evaluate our model's ability to understand cybersecurity content, facilitated by domain knowledge encoding, and to generate responses to cyber risk identification questions, enabled by SFT and RLHF training techniques. The original GPT-2 model, without training on pretrain data, serves as the baseline.

Firstly, the ability of understanding cybersecurity content is

Table 6
Cyber risk list from [7].

Phase (key1)	Potential threats (key2)	Potential vulnerabilities (key3)
Initiation	Data theft	Unsecured network transfer and cloud storage applications
Design	Proprietary information stolen	Unpatched software
Construction & Procurement	Performance degradation, physical damage	Excessive usage, fabricated chips
Commissioning	Data tampering, actuation tampering	Compromised dashboard and sensor
Operation & Maintenance	Spying, deliberate destruction	Chip insertion
Renovation & End of life	Data retrieval	Disposed sensors and equipment

evaluated [50], which can demonstrate the extent of cybersecurity knowledge the models have acquired gradually. Given phishing represents an emerging cyber risk in construction [40,65], we evaluated our models' performance in classifying phishing emails using the public Phishing Email Detection [66]. This dataset consists of 11,929 email texts, with 43.5% labeled as phishing attempts. We compared the performance of each model using accuracy, precision, recall, and F1 score [62].

Secondly, the ability of generating answers regarding cybersecurity is evaluated [50]. The literature [7] was selected as the benchmark because, to the best of the authors' knowledge, it is the only related work that offers a relatively more comprehensive list of cyber risks for construction projects. The findings of work [7] are presented in Table 6, from which the questions were formulated as 'Cyber risks in the {key1} phase include'. Each reference answer was structured as 'Cyber risks in the {key1} phase include {key2} resulting from {key3}.' Two sets of metrics were employed: (1) BERTScore [67], encompassing Precision (semantic similarity between generated and reference answers), Recall (coverage of reference answer's words by the generated answer), and F1 Score (the harmonic mean of Precision and Recall) for a balanced assessment of answer similarity in both wording and context. (2) GPT-4 [14] is requested to score from 0 to 100 based on Content (relevance, accuracy, and completeness), Clarity and Coherence (organization and readability), Specificity (detail and precision), and Risk Coverage (comprehensiveness of identified cyber risks) [68]. Priority is given to Content and Risk Coverage, with the assigned weights being 0.3, 0.2, 0.2, and 0.3, respectively.

3.5.2. Evaluating identified cyber risks

To achieve a comprehensive validation on the identified cyber risks, three sources are drawn: the selected benchmark [7], assessments by the two experts abovementioned, and assessments by GPT-4 [14]. To ensure a thorough and unbiased assessment of the experts and GPT-4, we designed their assessment process strategically:

- **Diverse Expertise Backgrounds:** Experts from different domains were invited to provide a broad spectrum of insights. The first expert, a cybersecurity specialist from the U.S., brought five years of industry experience. The second expert, a construction domain specialist from China, contributed six years of experience in the construction sector.
- **Multidimensional Judgement Criteria:** Assessors were tasked with evaluating each cyber risk from three attributes: its relevance to the construction phase, its likelihood of occurrence, and its potential impact on the project. Assessors evaluated each attribute using a 5-point Likert scale [60–62], where 1 signifies strong disagreement with the presence or significance of the attribute, and 5 indicates strong agreement.
- **Independent Assessment:** Assessments of the experts and GPT-4 were conducted independently. This approach ensured that each assessor evaluated the cyber risk list without prior knowledge of the other's opinions or ratings, thereby reducing the likelihood of conformity bias and encouraging impartial judgments.
- **Round Robin Evaluation:** We implemented a Round Robin evaluation method [69], dividing the assessment processes into six rounds, with each round corresponding to a project phase. Each assessor dedicated two days to the assessment, reviewing three project phases per day. Each phase was reviewed twice within the same day, and the average of the two assessments was taken to ensure an unbiased evaluation. This structured approach helped prevent assessment fatigue and ensured that each risk received adequate scrutiny.

Four dimensions of evaluation results are presented: (1) A detailed comparison with the benchmark literature [7]; (2) Qualitative feedback from two experts; (3) Analysis of the relevance of the identified cyber risks; and (4) Statistical analyses, aiming to verify the validity and consistency of likelihood assessments among all assessors. This involves comparing descriptive statistics to ensure likelihood assessments aligned with reality, utilizing the Friedman test [70] to ascertain the consistency of assessment criteria among all assessors holistically, employing the Wilcoxon Signed-Rank test [71] to determine if the assessment criteria between any two assessors are consistent, and applying the Spearman Rank Correlation test [72] to examine the consistency of risk prioritizations between any two assessors. These tests were chosen because they are non-parametric and do not assume a normal distribution, making them suitable for processing ordinal data, such as the risk likelihood levels. Additionally, they do not require a large sample size, rendering them appropriate for our analyses and providing reliable insights. This comprehensive set of results aims to further validate the effectiveness of our RL model and the applicability of the identified cyber risks. The details of implementations and results are presented in Section 4.3.

4. Implementations and results

4.1. Model training and selection

The training sessions were conducted using the PyTorch computation framework with a fixed random seed to ensure replicability. Additionally, we employed the default seed value of 42 for dataset splitting via the Scikit-learn package. The configuration details for the four stages are outlined in Table 7. The results are as follows:

(1) **Base model for stage 1.** Table 8 shows the training outcomes for the base models. Model selection considers the balance of

Table 7
Training details of the four stages.

Model type	Dataset split (train/test)	Epoch	Loss	Optimizer	Initial learning rate	Batch size
Base	95% / 5%	20	Cross-entropy loss using teacher forcing [20]	AdamW with scheduler	5e-5	32
SFT	80% / 20%	20	Cross-entropy loss using teacher forcing [20]			
Reward	80% / 20%	100	MSE loss			
RL	80% / 20%	20	Designed loss in Eq. (4)			

Table 8
Training results of base models (the first 10 epochs).

Model	Trainable parameters (million)	Training time (hours)	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8	E_9	E_{10}
GPT-2	124	8.5	3.711	3.634	3.610	3.603	3.608	3.622	3.646	3.659	3.677	3.683
BERT-LM	109	9	3.861	3.728	3.671	3.662	3.676	3.698	3.728	3.754	3.767	3.779
T5-LM	113 (Decoder only)	19	3.713	3.615	3.556	3.517	3.494	3.478	3.467	3.460	3.456	3.454

effectiveness and efficiency on the test set. All models showed similar test performance, but GPT-2 and BERT-LM were more time-efficient than T5-LM, resulting in T5-LM's exclusion. GPT-2 consistently surpassed BERT-LM, peaking at 3.603 after 4 epochs, and was chosen as the final base model for its sufficient encoding of construction cybersecurity knowledge.

(2) **SFT model for stage 2.** The training and test losses were computed solely on the answer part. The SFT model selection was based on its performance on the test set, as were the reward model and the RL model. The SFT model's peak performance on the test set was achieved at Epoch 12, with a loss of 0.015.

(3) **Reward model for stage 3.** The reward model achieved the lowest test loss at Epoch 69, exhibiting a prediction bias of ±6.5, which is considered acceptable and indicates reliable scoring predictions.

(4) **RL model for stage 4.** We experimented 32 combinations of weights for the three loss terms in Eq. (4), ensuring each weight is non-zero, a single digit, and sums to 1. For each combination, the model underwent training illustrated in Table 7. The model achieved its best test loss of 3.22 in Epoch 8 with a weight combination of (0.5, 0.2, 0.3). The higher weight assigned to the first term (the reward term) demonstrates the effectiveness of our reward model in improving the model's generalization ability of answering unseen questions.

4.2. Evaluating model progress

4.2.1. Progress in understanding cybersecurity content

Implementing the method in Section 3.5.1, the training was executed with a 5:1 ratio for splitting the training and testing data from the Phishing Email Detection dataset [66], spanning across 10 epochs. Fig. 4 delineates the metric improvements for each model, where a "*" symbol denotes the peak performance of a model in a specific metric. Impressively, all models recorded scores exceeding 0.9 across a variety of metrics, each surpassing the baseline. This achievement underscores the enhanced cybersecurity comprehension of our models, a direct

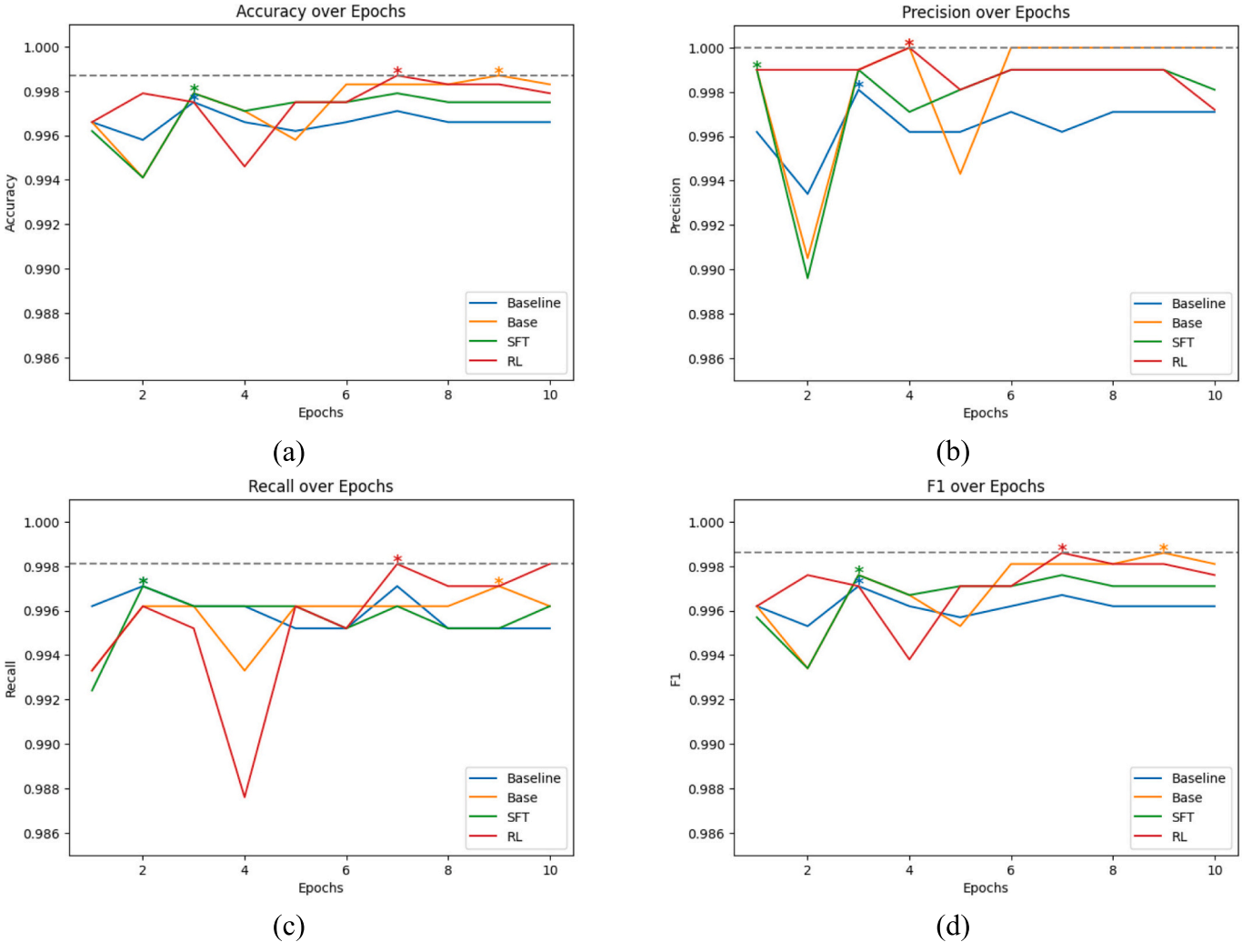


Fig. 4. Results on phishing text classification.

Table 9
Detailed evaluation results of answer generation ability.

Prompt & reference no.	Model	Precision	Recall	F1	GPT-4 (divided by 100)
	Baseline	0.4677	0.599	0.5253	0.0927
	Base	0.5940	0.7015	0.6433	0.6403
	SFT	0.5726	0.7107	0.6342	0.6986
1	RL	0.5734	0.7229	0.6354	0.7549
	Baseline	0.5646	0.635	0.5977	0.3184
	Base	0.4588	0.722	0.5611	0.2447
	SFT	0.5186	0.6933	0.5933	0.3869
2	RL	0.5308	0.7384	0.6004	0.5388
	Baseline	0.5123	0.6594	0.5766	0.2162
	Base	0.5473	0.7059	0.6165	0.6027
	SFT	0.5647	0.7274	0.6358	0.6997
3	RL	0.5855	0.7276	0.6454	0.7626
	Baseline	0.5584	0.5781	0.5681	0.2344
	Base	0.5873	0.5966	0.5919	0.5634
	SFT	0.5746	0.6738	0.6203	0.8568
4	RL	0.6205	0.6872	0.6411	0.7483
	Baseline	0.5201	0.6964	0.5954	0.3093
	Base	0.5447	0.7506	0.6313	0.6628
	SFT	0.5778	0.7494	0.6525	0.7468
5	RL	0.5691	0.7541	0.6487	0.7257
	Baseline	0.5475	0.726	0.6242	0.1123
	Base	0.5585	0.7599	0.6438	0.2533
	SFT	0.5608	0.7505	0.6419	0.6897
6	RL	0.5543	0.7271	0.6290	0.6940

Table 10
Answer generation evaluation results averaged across phases.

Model	Precision	Recall	F1	GPT-4	Average
Baseline	0.5284	0.649	0.5812	0.2139	0.4931
Base	0.5484	0.7061	0.6146	0.4945	0.5909
SFT	0.5615	0.7175	0.6297	0.6798	0.6471
RL	**0.5723**	**0.7262**	**0.6333**	**0.7041**	**0.6590**

consequence of utilizing our specially curated pretraining dataset. The culmination of this progression is observed with the RL model, which surpasses all others in every metric.

4.2.2. Progress in answering cybersecurity questions

Implementing the method in Section 3.5.1, Table 9 details the performance results, while Table 10 summarizes the average performance across all prompts. Generally, the models show improved performance on each metric, indicating a better ability to generate answers. Analyzing BERTScore metrics, the RL model surpasses others in all aspects. The RL model's higher precision and recall values indicate that its answers include relevant sections and cover a larger proportion of the reference, with the F1 score—the harmonic mean of both—highlighting its superior performance. The low baseline score (0.2139) by GPT-4 suggests its generated answer partially resembles the reference (from the moderate BERTScore) but lacks overall quality. Conversely, the RL model's high score (0.7041) indicates its answers are not only more similar to the reference but also meet the high quality of our scoring criteria. The average score, considering all metrics, reveals the RL model as the top performer (0.6590), indicating it is the most effective in generating satisfying answers, outperforming the baseline by 33.64%.

In summary, the model is progressing in both its ability to understand cybersecurity and generate answers. This progression verifies the efficacy of our training strategy pathway, thus laying the foundation for our RL model's effectiveness in performing the task of cyber risk identification.

4.3. Evaluating identified cyber risks

Following the methods in Sections 3.5 and 3.5.2, the identified cyber risks and likelihoods given by RL model are shown in Table 11, and the evaluation results are presented in following sections.

4.3.1. Superiority over the benchmark

Compared to the results of the work [7] (shown in Table 6), our language model and results demonstrate superiority in comprehensiveness, adaptability and speed, as shown in Table 12.

4.3.2. Positive feedback from experts

The cyber risk checklist received strong endorsements from both experts. The cybersecurity expert praised its comprehensiveness and relevance, highlighting its utility in segmenting risks by project phase and its role in enhancing awareness and prioritization of cybersecurity efforts within the construction industry. He recommended the checklist as an essential tool for companies aiming to bolster their cybersecurity posture. Similarly, the construction domain expert recognized the checklist's value in raising awareness of cyber threats and vulnerabilities, endorsing its broad coverage and foundational role in advancing cybersecurity practices. He advocated for a proactive cybersecurity approach guided by the checklist, suggesting collaboration with cybersecurity professionals to refine and adapt strategies to the dynamic cyber threat environment.

4.3.3. High relevance of identified risks

Table 13 displays the percentages of relevance levels by all assessors. It reveals that none of the risk items were rated with a relevance level below 3, with level 5 being the most common rating for both. This suggests that all the identified cyber risks on the list are considered relevant, demonstrating the effectiveness of our language models in identifying cyber risks.

4.3.4. Valid and consistent assessments

Following Section 3.5.2, statistical analyses were performed to check the validity and consistency of all assessors.

(1) **The assessments are demonstrated to be valid.** Looking at the descriptive statistics shown in Table 14, the variation in mean values across different phases for each assessor reflects the distinct cybersecurity challenges inherent to each phase, underscoring the necessity of the difference for phase-specific assessment criteria. Furthermore, the fluctuation in likelihoods assigned by each assessor within a phase, evidenced by the standard deviation, confirms that assessors recognize the uniqueness of each cyber risk. These variations add validity to the assessment criteria of all assessors, including our RL model.

(2) **The assessments are consistent among assessors overall.** We conducted phase-specific Friedman test [70], which is suitable for ordinal datasets (in our case, levels) and does not require minimum sample size, to check for overall significant differences within each phase in assessors' criteria. Because the likelihoods for all assessors are identical for the Construction & Procurement phase, the test was omitted for this phase. The null hypothesis (H_0) for every other phase is set to "There is no statistically significant difference in the assessments". The test results are presented in Table 14. It is evident that the *p*-value for these phases exceeds the alpha level of 0.05, so we fail to reject the null hypothesis H_0 for each phase. This suggests that the assessors are holistically adhering to the same criteria of phase-specific assessment, proving our RL model's effectiveness.

(3) **The assessments are consistent between any two assessors.** To make sure there is no significant differences between pairs of assessors that might have been overlooked by the Friedman test, we performed the pairwise Wilcoxon Signed-Rank test [71] as a post-hoc analysis. The null hypothesis (H_0) for each assessor pair in each phase is set to "There is no statistically significant difference in the assessments". The results are presented in Table 15.

Table 11
Cyber risk identification checklist with risk assessment results.

Phase	Cyber risks (threats and vulnerabilities)	Potential consequence	Expert 1			Expert 2			GPT-4			RL model	Ave L	Ave I	Risk value	Risk level*
			R	L	I	R	L	I	R	L	I	L				
Initiation	Weak identity and access management;	Lead to unauthorized access to systems and data, increasing the risk of data breaches, fraud, and compliance violations.	4	5	5	4	5	5	4	4	4	5	4.75	4.67	22.17	VH
	Unauthorized access to bidding documents and information;	Lead to the leakage of sensitive project details, giving competitors unfair advantages and potentially resulting in financial losses.	5	3	5	4	3	4	5	3	5	2	2.75	4.67	12.83	M
	Insecure communication channels;	Expose the transmission of critical and confidential data to interception by cyber attackers, risking the confidentiality and integrity of sensitive information.	5	2	4	5	2	5	4	2	5	2	2.00	4.67	9.33	L
	Weak permission controls for the use of assets and systems;	Allow unauthorized access and misuse of resources, leading to data leaks, system disruptions, and compromised security protocols.	4	2	4	4	1	5	4	2	4	2	1.75	4.33	7.58	L
	Insufficient data encryption for bidding documents and plans;	Can make it easier for unauthorized individuals to intercept and decipher confidential information, compromising the integrity and security of the project.	5	1	5	5	1	4	5	1	5	1	1.00	4.67	4.67	VL
	Phishing attacks targeting personnel of construction companies;	Result in unauthorized access to corporate systems and sensitive data, leading to data breaches and financial fraud.	3	1	3	4	1	4	4	1	5	1	1.00	4.00	4.00	VL
	Alteration of configuration data and/or information associated with digital facility operations;	Undermine the integrity of building systems, leading to operational inefficiencies and potential safety hazards.	5	5	5	4	5	5	5	5	5	5	5.00	5.00	25.00	VH
	Use of outdated versions of BIM and other design tools;	Leaving systems open to exploits and data breaches, and potentially leading to design flaws.	4	4	4	5	4	5	4	4	5	4	4.00	4.67	18.67	H
	Failure to update software with patches;	Serve as a gateway for cyberattacks, exposing systems to vulnerabilities and compromising the security of data and operations.	4	3	4	5	3	4	5	3	5	3	3.00	4.33	13.00	M
	Potential attempts to exploit vulnerabilities to access and erase data;	Disrupt operations, result in the loss of critical information, and significantly impact project timelines and costs.	5	3	5	4	2	4	5	2	4	2	2.25	4.33	9.75	L
Design	Lack of regulation in the data sharing process;	Lead to the inadvertent disclosure of sensitive information, violating privacy laws and undermining stakeholder trust.	5	2	4	5	3	3	5	3	3	3	2.75	3.33	9.17	L
	Unauthorized third-party access and utilization of confidential and/or proprietary information;	Lead to intellectual property theft, competitive disadvantage, and significant financial losses.	5	2	5	5	1	5	5	2	5	1	1.50	5.00	7.50	L
	Unauthorized access to design and construction information (e.g., access control card, door position, CCTV system, PINs, etc.);	Compromise the security of a facility, making it susceptible to physical breaches and endangering the safety of occupants.	5	2	5	5	1	5	5	1	4	1	1.25	4.67	5.83	L
	Fabrication of documents, images, and information pertaining to design and usage;	Result in non-compliance with regulations, jeopardize safety standards, and lead to financial and reputational damage.	5	1	5	5	1	5	4	1	5	1	1.00	5.00	5.00	VL
Construction & Procurement	Unauthorized access to information on design and construction;	Lead to the exposure of proprietary designs and construction methodologies, jeopardizing competitive	5	5	5	4	5	4	5	5	4	5	5.00	4.33	21.67	VH

(*continued on next page*)

Table 11 (*continued*)

Phase	Cyber risks (threats and vulnerabilities)	Potential consequence	Expert 1			Expert 2			GPT-4			RL model	Ave L	Ave I	Risk value	Risk level*
			R	L	I	R	L	I	R	L	I	L				
		advantage and project integrity.														
	Difficulty accessing data or information about the contractor, operator, and maintenance owner;	Hinder project coordination, delay construction timelines, and impact overall project management efficiency.	4	2	4	5	2	5	4	2	5	2	2.00	4.67	9.33	L
	Interference with the operation or integrity of devices (also referred to as data theft) and manipulation of processes (IoT);	Compromise system functionality, endanger safety, and lead to unauthorized control or data breaches.	5	1	5	4	1	5	3	1	5	1	1.00	5.00	5.00	VL
	Manipulation of construction delivery services and other construction-related systems through cyber manipulation;	Disrupt project timelines, increase costs, and compromise safety protocols.	5	1	5	5	1	4	5	1	5	1	1.00	4.67	4.67	VL
	Limited access controls and identity verification	Allow unauthorized individuals to access sensitive systems and data, increasing the risk of malicious activities and compromising the commissioning process.	5	5	5	5	5	4	5	5	4	5	5.00	4.33	21.67	VH
	Unauthorized access to commissioning data and systems	Compromise the integrity of the commissioning process, potentially leading to operational failures or safety hazards in the final infrastructure.	5	5	5	5	4	5	4	3	4	4	4.00	4.67	18.67	H
	Infiltration of malware into the commissioning systems	Disrupt critical operations, damage systems, and lead to costly delays and repairs, undermining the reliability of the commissioned asset.	5	2	5	5	2	5	5	1	5	2	1.75	5.00	8.75	L
Commissioning	Social engineering attacks targeting commissioning personnel	Result in unauthorized access to systems and data, endangering the security of the commissioning operations and potentially leading to data breaches.	5	2	5	4	1	5	4	1	5	1	1.25	5.00	6.25	L
	Insufficient cybersecurity measures for IoT devices utilized in commissioning	Leave these devices vulnerable to attacks, compromising the functionality and security of critical systems involved in commissioning activities.	5	1	5	5	1	5	5	1	5	1	1.00	5.00	5.00	VL
	Vulnerable communication channels between commissioning teams and systems	Expose sensitive data to interception and manipulation, risking the confidentiality and integrity of the commissioning process.	5	1	4	3	1	3	4	1	2	1	1.00	3.00	3.00	VL
	Leakage of or interference with the critical asset information;	Undermine the security and integrity of essential services, leading to operational disruptions and potential safety hazards.	5	5	5	5	5	5	5	5	5	4	4.75	5.00	23.75	VH
	Unauthorized access to the design and construction data/information;	Lead to the exposure of proprietary or sensitive project details, potentially resulting in competitive harm and financial loss.	4	5	4	4	5	5	5	5	5	5	5.00	4.67	23.33	VH
Operation & Maintenance	Leakage of or interference with the construction model, the operation and maintenance plan of the assets;	Disrupt operational efficiency, compromise safety protocols, and lead to increased maintenance costs.	5	2	5	4	3	5	5	3	5	3	2.75	5.00	13.75	M
	Malicious actors who obtain access to a building's systems and data, such as keystroke loggers, code-based logic bomb threats, file downloads, etc.	Introduce risks such as espionage, sabotage through logic bombs, and unauthorized access to sensitive information, compromising security and privacy.	5	1	5	4	1	4	4	2	5	2	1.50	4.67	7.00	L
	Attacks on the operational phase, which may lead to	Cause physical damage to infrastructure, result in the	5	1	5	5	1	4	5	1	4	1	1.00	4.33	4.33	VL

(*continued on next page*)

Table 11 (*continued*)

Phase	Cyber risks (threats and vulnerabilities)	Potential consequence	Expert 1			Expert 2			GPT-4			RL model	Ave L	Ave I	Risk value	Risk level*
			R	L	I	R	L	I	R	L	I	L				
Renovation & End of life	physical damage, theft of intellectual property, and damage to third-party assets;	theft of intellectual property, and lead to financial liabilities due to damage to third-party assets.														
	Inadequate identity and access management for demolition-related systems.	Allow unauthorized access to critical controls and information, increasing the risk of sabotage, data theft, and manipulation of the demolition process.	5	5	5	5	5	5	5	4	4	5	4.75	4.67	22.17	VH
	Unauthorized access to demolition plans and schedules;	Lead to premature disclosures or alterations, potentially endangering workers and the public by compromising the planned safety measures.	5	4	5	5	3	5	5	4	5	3	3.50	5.00	17.50	H
	Disclosure or theft of sensitive demolition-related information;	Expose strategic or competitive details, leading to financial losses or unauthorized access to secured sites, compromising safety and security protocols.	5	3	5	5	1	5	4	2	4	2	2.00	4.67	9.33	L
	Insecure communication and data transmission among stakeholders;	Result in the interception or manipulation of sensitive data, undermining the integrity and confidentiality of the demolition operation.	5	2	4	5	2	5	5	2	5	2	2.00	4.67	9.33	L
	Social engineering attacks aimed at demolition personnel;	Trick individuals into divulging confidential information or granting access to restricted systems, jeopardizing the security of the demolition operation.	5	2	4	5	2	3	5	3	4	2	2.25	3.67	8.25	L
	Interference with demolition-related systems and safety controls;	Cause malfunctions or failures, leading to accidents, injuries, or uncontrolled collapses, significantly increasing risk to human life and surrounding properties.	5	1	5	5	2	4	5	2	3	2	1.75	4.00	7.00	L
	Unpatched or outdated software utilized in the demolition process;	Contain vulnerabilities that cyber attackers could exploit, potentially leading to system failures, data breaches, or unauthorized control of demolition activities.	4	1	4	3	1	4	4	2	4	1	1.25	4.00	5.00	VL

* Notation: R: Relevance; L: Likelihood; I: Impact; VH: Very High; H: High; M: Medium; L: Low; VL: Very Low.

It shows that *p*-values across all phases are greater than the alpha level of 0.008, which has been adjusted using the Bonferroni correction [71] to mitigate the increased risk of errors from multiple pairwise comparisons, so we fail to reject H_0 for each assessor pair, indicating no significant difference in assessment criteria between any two assessors. This further confirms the consistency of the assessors, proving our RL model's effectiveness.

(4) **The risk prioritizations are consistent among assessors.** Spearman Rank Correlation test [72] was adopted to determine whether risk prioritization within each phase was consistent across assessors, as indicated by their correlations. The null hypothesis (H_0) for each pair of assessors in each phase was set as "There is no statistically significant correlation in the assessments". The test results are presented in Fig. 5, where the right upper triangle of each heatmap contains the *p*-values, while the left lower triangle contains the correlation coefficients. All coefficients are greater than 0.5, indicating a strong correlation among all assessor pairs. An exception was Expert 2 - Expert 3 in the Renovation & End of Life phase, which had a p-value of 0.123. However, nearly all other p-values were significantly below the alpha level of 0.05, so we rejected the H_0 for these pairs, demonstrating statistically significant correlations. These high correlations indicate the consistency in risk prioritizations among assessors, further validating the accuracy across assessors and proving the effectiveness of our RL model.

In summary, the results from the benchmark comparison, expert and GPT-4 assessments, and statistical analyses further validate the effectiveness of our RL model and the applicability of the identified risks. Table 11 presents the identified risks for each phase along with the consequences elicited from experts, ranked by their risk values.

5. Discussions

5.1. The applicability of the checklist

Two main applications of the prioritized checklist are proposed: (1) It can serve as a new benchmark, which project managers can refer to for formulating proactive and preventive measures for their projects, focusing on the most significant risks. The process of developing these measures can involve collaboration between project managers, IT, and security teams. Such prioritization helps in effectively preventing high

Table 12
Comparison with the selected benchmark [7].

Aspect	Sub-aspect	Benchmark [7]	Our work
Comprehensiveness	List of Cyber Risks	Not comprehensive, providing one or two identified risks for each phase	Comprehensive, providing detailed list across different project phases and ensuring broad applicability to construction projects
	Risk Prioritization	Not specified	Specified ranking of cyber risks (Table 11), aiding strategic resource allocation for risk prevention and mitigation
	Scenario Coverage	Not wide, may missing specific incident scenarios	Wide, covering a broad spectrum of incident scenarios and can be updated regularly
Adaptability	Framework Nature	Static, requiring complete restart for updating the list	Dynamically updateable, supporting model self-training for ongoing adaptation to new data
	Response to Cybersecurity Landscape Changes	Limited, necessitating expert intervention and framework restart for updates	Capable, efficiently aligning with changes, allowing periodic updates without training the model from scratch
	Updating Process	Complicated, requiring a group of experts for framework adjustments and restarting	Simple, allowing automatically updating with new data, minimizing the need for expert intervention
Speed	Time to Identify Risks	Time-consuming, requiring days to weeks per cycle, dependent on manual processes	Time-efficient, although initially taking around 87 h for training and fine-tuning (excluding time for experimentations), but allowing for rapid adaptation to new data within hours
	Human Intervention	High, reliant on manual processes and expert discussions	Minimal, primarily for oversight, with the language model handling the bulk of processing
	Inference Speed	Not applicable, requiring restarting the whole process	Fast, requiring only seconds for answering a question, enabling swift risk re-identification and response to evolving threats

risks while ensuring more efficient resource allocation. (2) Risk analysts can use the checklist for in-depth risk analyses on specific projects, focusing first on the most significant risks. This process may involve identifying risk factors, performing quantitative risk assessments, and pinpointing critical risk factors, which should be addressed with priority.

In addition to the two primary applications, the checklist serves as a vital tool for IT and cybersecurity teams to evaluate and enhance existing security protocols. By identifying specific vulnerabilities, these teams can implement more robust security measures tailored to the unique needs of construction projects. Furthermore, stakeholders, including contractors and vendors, can use the checklist to verify that their systems and communication methods meet the necessary security standards. Moreover, the checklist can also play a crucial role in education and training, enabling personnel to understand the nature of common cyber risks, understand their potential impact, and recognize the importance of adhering to established security protocols. As the checklist can be regularly updated, people will have access to the latest information on the cybersecurity landscape, ensuring they remain informed about current trends.

5.2. Risk mitigation recommendations

Among the “High” and “Very High” risks, the primary concerns in construction projects can be summarized as weak identity and access management, unauthorized access across various project stages, falsification of critical data, reliance on outdated software tools, and leakage of crucial asset information. These areas require attention. Additionally, it is evident that cyber risks can stem from both the IT sector and the construction sector, affecting the management and operational aspects of a project. This intersectionality of risk sources underscores the necessity for resolutions that involve collaboration between IT and construction professionals. By collaborating, these experts can ensure that cybersecurity measures are not only in the technical aspect but also encompass the optimization of management and operational guidelines, all of which should be practically applicable within the unique context of construction projects. The integration of IT security practices with construction management processes is integral in the project lifecycle.

We recognize that the cyber risks identified are high-level and broadly applicable to diverse projects. Considering the variations among individual projects, each has different exposures to these risks, attributed to the unique risk factors inherent to each project. Therefore, it is crucial to closely examine the specific factors contributing to each risk for individual projects, so as to formulate efficient and targeted risk mitigation strategies at the project level [73]. Key factors for investigation include various aspects of construction projects: general project information, project structure, IT-related factors, OT-related factors, and the human and management aspects [73]. Such a study will not only enable us to quantify the risk level specific to a construction project but also evaluate the effectiveness of risk mitigation strategies by directly addressing these critical factors.

5.3. The prospect of the language model

In Section 4.2.2, the quantitative evaluation of the models’ progress has shown that the RL model excels in generating responses to cybersecurity queries. Providing an in-depth qualitative illustration, Table 16 compares the responses of our RL model and the baseline model (original GPT-2) to five varied prompts, all concerning the definition of phishing. The RL model consistently delivers relevant and informative responses, effectively explaining the phishing concept regardless of the phrasing differences. In contrast, the baseline model frequently yields answers that are irrelevant to the prompts. These evaluations demonstrate the RL model’s superior capability in handling cybersecurity queries, underscoring its potential as an effective cybersecurity consultant for construction personnel, particularly for those with limited knowledge.

Table 13
Percentage of relevance levels.

Expert 1			Expert 2			GPT-4		
Level 3	Level 4	Level 5	Level 3	Level 4	Level 5	Level 3	Level 4	Level 5
2.78%	19.44%	75.00%	5.56%	25.00%	69.44%	2.78%	30.56%	66.67%

Table 14
Descriptive statistics and Friedman test results.

Phase	RL model (Mean ± SD)	Expert 1 (Mean ± SD)	Expert 2 (Mean ± SD)	GPT-4 (Mean ± SD)	Test statistic	*P* value
Initiation	2.167 ± 1.344	2.333 ± 1.374	2.167 ± 1.462	2.167 ± 1.067	1.000	0.801
Design	2.500 ± 1.414	2.750 ± 1.199	2.500 ± 1.414	2.625 ± 1.317	2.538	0.468
Construction &Procurement	2.250 ± 1.639	2.250 ± 1.639	2.250 ± 1.639	2.250 ± 1.639	–	–
Commissioning	2.333 ± 1.599	2.667 ± 1.700	2.333 ± 1.599	2.000 ± 1.528	7.000	0.072
Operation & Maintenance	3.000 ± 1.414	2.800 ± 1.833	3.000 ± 1.789	3.200 ± 1.600	2.400	0.494
Renovation & End of life	2.429 ± 1.178	2.571 ± 1.400	2.286 ± 1.278	2.714 ± 0.881	2.415	0.491

Table 15
Wilcoxon signed-rank test results.

Initiation	RL model	Expert 1	Expert 2	GPT-4
RL model	–	1.000	1.000	1.000
Expert 1	0.000	–	1.000	1.000
Expert 2	1.500	0.000	–	1.000
GPT-4	1.500	0.000	1.500	–
Design	RL model	Expert 1	Expert 2	GPT-4
RL model	–	0.375	1.000	1.000
Expert 1	2.500	–	0.625	0.750
Expert 2	2.500	2.500	–	1.000
GPT-4	0.000	2.000	0.000	–
Construction & Procurement	RL model	Expert 1	Expert 2	GPT-4
RL model	–	1.000	1.000	1.000
Expert 1	0.000	–	1.000	1.000
Expert 2	0.000	0.000	–	1.000
GPT-4	0.000	0.000	0.000	–
Commissioning	RL model	Expert 1	Expert 2	GPT-4
RL model	–	0.500	1.000	0.500
Expert 1	0.000	–	0.500	0.250
Expert 2	0.000	0.000	–	0.500
GPT-4	0.000	0.000	0.000	–
Operation & Maintenance	RL model	Expert 1	Expert 2	GPT-4
RL model	–	0.750	1.000	1.000
Expert 1	2.000	–	1.000	0.500
Expert 2	1.500	0.000	–	1.000
GPT-4	0.000	0.000	0.000	–
Renovation & End of life	RL model	Expert 1	Expert 2	GPT-4
RL model	–	0.750	1.000	0.375
Expert 1	2.000	–	0.750	0.813
Expert 2	0.000	1.500	–	0.313
GPT-4	2.500	6.000	3.000	–

Note: Bonferroni correction: $\alpha = 0.008$. Underlined values represent the Wilcoxon signed-rank test statistic; all other values are p-values.

This model could be integrated into websites or apps, making it accessible to a wider audience within the construction industry, including contractors, project managers, safety officers, and stakeholders. In our experiment, the inference time for the model to generate answers is around 2 s per prompt. This response time is reasonable and acceptable for deployment in real-world applications. Additionally, if the model is updated with new data on cybersecurity in construction projects—recommended after every 10,000 new sentences—the estimated training time is around 1.5 h. This duration is manageable and also considered acceptable for ongoing maintenance. With regular updates, the model can provide project-specific risk identification and analysis, dynamic risk identification and analysis, real-time risk monitoring, solution recommendations, and customized cybersecurity training for stakeholders. Update can also focus on enlarging the model size, aiming to develop a construction cybersecurity-specific LLM that is 10^4 times larger, comparable to GPT-4 [14]. Its capacity for human-like interactions will improve the system's ability to comprehend and process a variety of inquiries from different users, thus enabling the model to grasp users' intentions more accurately. The model's responses will also be more comprehensible to users. This interactive feature can effectively assist individuals of diverse backgrounds and educational levels, increasing its utility for widespread industry application.

5.4. Enhancing dataset for model upgrade

Regarding a larger, more capable model, there are several considerations to keep in mind. Firstly, the dataset used to train our base model, compiled from online sources, needs update and expansion. Considering the vast amount of text generated daily, new data should be collected periodically and organized, then integrated with existing datasets to enhance the model's understanding of construction cybersecurity. This approach will also keep the model current with cybersecurity trends. The frequency of updates will depend on available computing resources and the current demand. Secondly, increasing the diversity of the SFT dataset could be advantageous. Our current SFT question-answering dataset primarily includes simpler questions that focus on definitions, significance, or impacts of specific entities. Future iterations could introduce more complex and varied questions, incorporating logic and reasoning tasks to challenge the model with more sophisticated assignments. For example, questions might require the model to evaluate and discuss the evolving cybersecurity status based on detailed project information, thus aligning the model more closely with project-specific issues. Thirdly, the criteria for scoring answers in the RM dataset can be more granular and comprehensive. This aspect is crucial as it significantly influences the reward model's understanding of what constitutes a 'high-quality' answer, which, in turn, affects the RL model's response quality. Future evaluations can consider a broader array of scoring rubric, including format, content, typographical errors, repetitions, and logical coherence.

5.5. Prompt formulation

Despite the powerful capabilities of language models, effectively communicating with them remains a challenge. The formulation of prompts crucially influences the quality of the generated answers. To address this, Section 3.4 details strategic methods for crafting prompts that elicit relevant and insightful responses from our developed language model, aiming to identify cyber risks across project phases. Four key considerations for prompt formulation are as follows:

(1) Ensuring Answer Relevance with Keywords: Keywords in the prompt should be straightforward to ensure that the model fully understands and generates relevant responses. In this study, the primary goal is to enable the model to provide insights into cyber risks across different project phases. Therefore, the prompt should include a set of keywords related to cyber risks and project phases to maintain relevance.
(2) Limit Prompt Wording Complexity: When working with smaller models, such as GPT-2 with around a million trainable parameters, the prompt should be concise and straightforward. This helps the model better understand and capture the essence of the prompt. In this study, the maximum number of words in our prompts is fewer than 15.

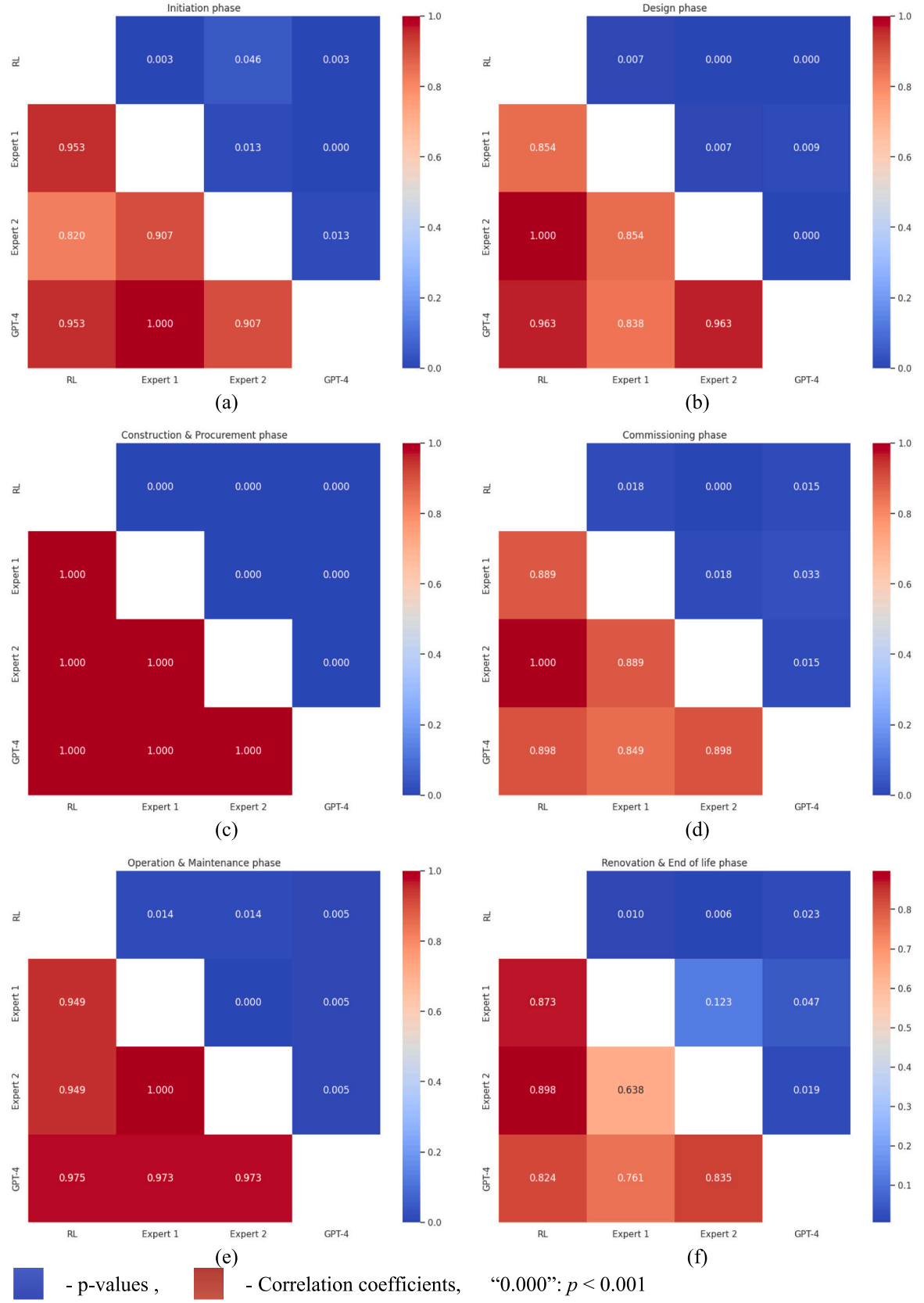


Fig. 5. Spearman’s rank correlation coefficient test results.

Table 16
Illustration of answers by the baseline and RL model.

Prompt	Baseline model	RL model
What is phishing?	We'll see what it looks like in the screenshots below, but before diving into the actual attack, please feel free to look at the phishing attacks in step 3.\n \nStep 2. Create a simple message or e-mail containing your password\n\nStep…	A phishing attack is a type of social engineering attack in which an attacker sends an unsolicited email or message, often disguised as a legitimate source, with the intent of tricking the recipient into divulging sensitive information.\nA phishing attack is a type of social…
How to define phishing?	The most common way for someone to use your e-mail address or your social media handle is to have them use it in the e-mail address or your social media handle and that address. This way, everyone uses your email address in order to spam you…	Phishing is a type of cyberattack in which attackers send fraudulent emails or messages to trick individuals into divulging sensitive information.\nPhishing is a type of cyber attack in which attackers send fraudulent emails or messages to trick individuals into divulging sensitive information.\nPh…
Can you explain phishing?	How do I get rid of it? (5:30)\n \nWhat are the main security features of phishing? (2:15)\n \nWhat are security flaws? (2:20)\n\nWhat does the system look like in real life…	Phishing is a type of cyber attack in which attackers send fraudulent emails or messages, often disguised as legitimate sources, with malicious attachments or links.\nA phishing attack can be conducted by attackers who view the emails or messages as legitimate sources. However, it can…
Discuss what phishing is.	Let's set up our system. The first thing you need to do is to do some common actions in your app:\n\nLaunch a new browser. In your app, go back to the menu bar (if using the built-in Chrome…	Phishing attacks, especially phishing attacks, can be conducted by attackers who view unsuspecting individuals or organizations as potential targets, using psychological manipulation techniques, or merely by sending fraudulent emails.\nPhishing attacks could be perpetrated…
Define phishing.	There are two types of phishers - those that make a mockery of your online service and those who make fun of you to ruin your online presence. Once you've discovered why your identity has been compromised, it's time to turn that around.\n \nA phishing attack…	Phishing is a type of cyber attack in which attackers send fraudulent emails or messages to trick individuals into divulging sensitive information or downloading malware. \nPhishing is a type of cyber attack in which attackers send fraudulent emails or messages to trick individuals into divulging sensitive information or…

(3) Impact of Keyword Differentiation: Language models may not excel at recognizing synonyms of various words. In this study, we consider incorporating different synonyms in one prompt to offset potential negative impacts. For example, one set of keywords includes various terms for cyber risk, while another set covers synonyms for project phases.
(4) Variation in Prompt Styles: The language model might be sensitive to specific sentence formats, so varying the format of the prompts can be beneficial. In Section 3.4, we employed five different styles for each prompt, including interrogative and imperative styles. These styles align with those used during the training process. Using diverse styles helps the model generate more comprehensive answers and avoids missing potential responses.

The aforementioned prompting strategies can improve interaction quality with language models and also serve as a guide for organizations. When customizing their own language models, companies can consider these strategies to formulate their own set of prompts for the tasks at hand.

5.6. Limitations and outlook

This study has limitations, including a dataset lacking project-specific information, limiting the model's ability to identify and assess unique project-specific cyber risks. Additionally, due to limited model size and computational resources, the risk identification process is only semi-automatic; while the model can generate content with risk identification results, it cannot fully automatically generate a checklist. Besides addressing the limitations, future efforts will focus on expanding our dataset and enhancing our model to match the capabilities of advanced models like Ernie Bot [13], GPT-4 [14] and Gemini [15]. We also plan to enable document analysis, including drawings, software codes, and schedules, to produce customized outputs such as figures, tables, and LaTeX code. Our final goal is to integrate this tool into web or mobile applications, offering an intelligent cybersecurity consultant accessible to various groups, especially those lacking cybersecurity expertise. This integration can achieve advanced functions, including project-specific risk analysis, dynamic risk identification and analysis, real-time risk monitoring, solution recommendations, customized cybersecurity training for stakeholders, etc.

6. Conclusions

A language model was developed in this paper to thoroughly identify cyber risks across project phases, which are applicable to diverse construction projects. Our model, trained on 61,841 sentences of construction cybersecurity textual data and enhanced by the adapted SFT and RLHF training techniques, shows expected improvement in understanding cybersecurity content and in answering cybersecurity questions. This positions our language model as suitable for identifying cyber risks across project phases. The cyber risk checklist, ranked by risk values, surpasses the existing literature, has received positive feedback from industry experts, and proves to be highly relevant. The risk likelihood assessments by our model are consistent with those of two experts and GPT-4. These results collectively validate the effectiveness of our model and the applicability of the identified cyber risks. This study discusses the applicability of the identified cyber-risk checklist, recommends high-level risk mitigation strategies, examines the potential of our developed language model, discusses in depth how to enhance the dataset for model improvement, and defines prompt formulation strategies for organizations.

Compared to previous studies and practices that mostly propose frameworks or methods reliant on manual effort and thus lack flexibility and efficiency, our developed language model offers a more comprehensive approach to cyber risk identification. It also enables dynamic identification in the future that ensures time efficiency. Moreover, two additional benefits are offered: (1) Customizability. The model can be fine-tuned with a company's own text corpus; fine-tuning requires only a few hours, and subsequent question-answering takes mere seconds per question. This capability facilitates efficient and tailored risk identification and analysis that considers the unique characteristics of each organization. (2) Intelligent cybersecurity consultant. The upgraded model can be used as an intelligent cybersecurity consultant deployed in construction companies' mobile or website applications, achieving various advanced functions and thus showing great potential for industry-wide utilization.

The academic contributions of this study include the development of a language model dedicated to construction cybersecurity, promoting interdisciplinary research in AI, cybersecurity, and construction management. It also bridges the gap in the comprehensive recognition of cyber risks across project phases, providing a new cyber risk checklist

benchmark. Practically, the cyber risk checklist is useful for diverse construction projects, particularly valuable in assisting project managers to check their cybersecurity status and formulate proactive and preventive cybersecurity measures against prioritized risk items. Additionally, risk analysts can leverage this checklist for in-depth risk analyses on specific construction projects, prioritizing their efforts by starting with the most critical risks in the list. Moreover, IT and cybersecurity teams, stakeholders and general personnel of construction companies can benefit as well. As the checklist can be regularly updated, people will have access to the latest information on the cybersecurity landscape, ensuring they remain informed about current trends.

Future work will focus on addressing the limitations of our dataset, which lacks project-specific information. Also, we aim to achieve a fully automated risk identification process by increasing the model size and utilizing more computational resources. Our long-term goal is to enhance the model to be comparable with large language models and integrate it into web and mobile applications for construction companies, enabling it to function as an intelligent cybersecurity consultant that provide various functions and accessible to a wide range of people.

CRediT authorship contribution statement

Dongchi Yao: Writing – original draft, Visualization, Validation, Methodology, Formal analysis, Data curation, Conceptualization. **Borja García de Soto:** Writing – review & editing, Supervision, Project administration, Investigation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported by the Center for CyberSecurity (CCS), funded by Tamkeen under the NYUAD Research Institute Award G1104 in collaboration with the NYUAD Center for Interacting Urban Networks (CITIES), funded by Tamkeen under the NYUAD Research Institute Award CG001. This work was also supported in part through the NYU IT High-Performance Computing resources, services, and staff expertise.

References

[1] B. García de Soto, I. Agustí-Juan, S. Joss, J. Hunhevicz, Implications of construction 4.0 to the workforce and organizational structures, Int. J. Constr. Manag. 22 (2) (Jan. 2022) 205–217, https://doi.org/10.1080/15623599.2019.1616414.
[2] H. Kayan, M. Nunes, O. Rana, P. Burnap, C. Perera, Cybersecurity of Industrial Cyber-Physical Systems: A Review, arXiv, Jan. 10, 2021. Accessed: Mar. 02, 2024. [Online]. Available: http://arxiv.org/abs/2101.03564.
[3] A. Parks, The ongoing risk of phishing in the construction industry, in: Construction Management Association of America, 2021. Accessed: Mar. 15, 2022. [Online]. Available: https://resources.infosecinstitute.com/topic/phishing-attacks-construction-industry/.
[4] T. Sawyer, J. Rubenstone, Construction cybercrime is on the rise, in: Expertise, Solution, and People for Success, 2019. Accessed: Apr. 23, 2021. [Online]. Available: https://www.cdp-inc.com/articles/construction-cybercrime-rise.
[5] C. Tunney, Ransomware Attack on Construction Company Raises Questions About Federal Contracts, CBC News, 2020. Accessed: Mar. 15, 2021. [Online]. Available: https://www.cbc.ca/news/politics/ransomware-bird-construction-military-1.5434308.
[6] R. Korman, Hoffman Construction Reports Hack of Self-Insured Health Plan Data, Engineering News-Record, 2021. Accessed: Mar. 15, 2021. [Online]. Available: https://www.enr.com/articles/51232-hoffman-construction-reports-hack-of-self-insured-health-plan-data.
[7] B. Mantha, B. García de Soto, R. Karri, Cyber security threat modeling in the AEC industry: an example for the commissioning of the built environment, Sustain. Cities Soc. 66 (Mar. 2021) 102682, https://doi.org/10.1016/j.scs.2020.102682.
[8] Ž. Turk, B. García de Soto, B.R.K. Mantha, A. Maciel, A. Georgescu, A systemic framework for addressing cybersecurity in construction, Autom. Constr. 133 (Jan. 2022) 103988, https://doi.org/10.1016/j.autcon.2021.103988.
[9] B. Zhong, X. Pan, P.E.D. Love, J. Sun, C. Tao, Hazard analysis: a deep learning and text mining framework for accident prevention, Adv. Eng. Inform. 46 (Oct. 2020) 101152, https://doi.org/10.1016/j.aei.2020.101152.
[10] A. Bittar, S. Velupillai, A. Roberts, R. Dutta, Using general-purpose sentiment lexicons for suicide risk assessment in electronic health records: corpus-based analysis, JMIR Med. Inform. 9 (4) (Apr. 2021) e22397, https://doi.org/10.2196/22397.
[11] M. Boholm, Å. Boholm, Risk identification: a corpus-assisted study of websites of government agencies, Risk Haz. Crisis Publ. Policy 11 (3) (Sep. 2020) 242–269, https://doi.org/10.1002/rhc3.12184.
[12] D. Yao, B. García de Soto, A corpus database for cybersecurity topic modeling in the construction industry, in: Presented at the 40th International Symposium on Automation and Robotics in Construction, Chennai, India, Jul. 2023, https://doi.org/10.22260/ISARC2023/0072.
[13] Baidu Inc, Introducing ERNIE 3.5: Baidu's Knowledge-Enhanced Foundation Model Takes a giant Leap Forward. http://research.baidu.com/Blog/index-view?id=185.
[14] OpenAI, et al., GPT-4 Technical Report, arXiv, Dec. 18, 2023. Accessed: Mar. 02, 2024. [Online]. Available: http://arxiv.org/abs/2303.08774.
[15] Gemini Team, et al., Gemini: A Family of Highly Capable Multimodal Models, arXiv, Dec. 18, 2023. Accessed: Mar. 02, 2024. [Online]. Available: http://arxiv.org/abs/2312.11805.
[16] W.X. Zhao, et al., A Survey of Large Language Models, arXiv, Nov. 24, 2023. Accessed: Mar. 02, 2024. [Online]. Available: http://arxiv.org/abs/2303.18223.
[17] B.R.K. Mantha, B. García de Soto, Cyber security challenges and vulnerability assessment in the construction industry, in: Proceedings of the Creative Construction Conference 2019, Budapest University of Technology and Economics, 2019, pp. 29–37, https://doi.org/10.3311/CCC2019-005.
[18] E.A. Parn, D. Edwards, Cyber threats confronting the digital built environment: common data environment vulnerabilities and blockchain deterrence, Eng. Constr. Archit. Manag. 26 (2) (Mar. 2019) 245–266, https://doi.org/10.1108/ECAM-03-2018-0101.
[19] N. Salami Pargoo, M. Ilbeigi, A scoping review for cybersecurity in the construction industry, J. Manag. Eng. 39 (2) (Mar. 2023) 03122003, https://doi.org/10.1061/JMENEA.MEENG-5034.
[20] D. Jurafsky, J.H. Martin, Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, 2 ed., Prentice Hall, Upper Saddle River, NJ, 2009 [Nachdr.]. (ISBN: 978-0-13-187321-6).
[21] A. Vaswani, et al., Attention is All You Need, arXiv, Aug. 01, 2023. Accessed: Mar. 02, 2024. [Online]. Available: http://arxiv.org/abs/1706.03762.
[22] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, Improving Language Understanding by Generative Pre-Training, Accessed: Mar. 02, 2024. [Online]. Available: https://openai.com/research/language-unsupervised, 2018.
[23] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding, arXiv, May 24, 2019. Accessed: Mar. 02, 2024. [Online]. Available: http://arxiv.org/abs/1810.04805.
[24] Radford Alec, Wu Jeffrey, Child Rewon, Luan David, Amodei Dario, Sutskever Ilya, Language Models are Unsupervised Multitask Learners, OpenAI blog, 2019. Accessed: Mar. 02, 2024. [Online]. Available: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.
[25] T.B. Brown, et al., Language Models are Few-Shot Learners, arXiv, Jul. 22, 2020. Accessed: Mar. 02, 2024. [Online]. Available: http://arxiv.org/abs/2005.14165.
[26] OpenAI, Introducing ChatGPT, OpenAI, 2022. Accessed: Mar. 02, 2024. [Online]. Available: https://openai.com/blog/chatgpt.
[27] A. Chowdhery, et al., PaLM: Scaling Language Modeling with Pathways, arXiv, Oct. 05, 2022. Accessed: Mar. 02, 2024. [Online]. Available: http://arxiv.org/abs/2204.02311.
[28] R. Thoppilan, et al., LaMDA: Language Models for Dialog Applications, arXiv, Feb. 10, 2022. Accessed: Mar. 02, 2024. [Online]. Available: http://arxiv.org/abs/2201.08239.
[29] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, O. Klimov, Proximal Policy Optimization Algorithms, arXiv, Aug. 28, 2017. Accessed: Mar. 02, 2024. [Online]. Available: http://arxiv.org/abs/1707.06347.
[30] R.S. Sutton, D. McAllester, S. Singh, Y. Mansour, Policy gradient methods for reinforcement learning with function approximation, in: Advances in Neural Information Processing Systems 12 (NIPS 1999), 1999, pp. 1057–1063.
[31] J.W. Rae, et al., Scaling Language Models: Methods, Analysis & Insights from Training Gopher, arXiv, Jan. 21, 2022. Accessed: Mar. 02, 2024. [Online]. Available: http://arxiv.org/abs/2112.11446.
[32] Y. Bai, et al., Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback, arXiv, Apr. 12, 2022. Accessed: Mar. 02, 2024. [Online]. Available: http://arxiv.org/abs/2204.05862.
[33] National Institute of Standards and Technology, Framework for Improving Critical Infrastructure Cybersecurity, Version 1.1, National Institute of Standards and Technology, Gaithersburg, MD, Apr. 2018, https://doi.org/10.6028/NIST.CSWP.04162018. NIST CSWP 04162018.
[34] G. D. P. Regulation, GDPR (General Data Protection Regulation), Intersoft Consulting, 2018. Accessed: May 16, 2023. [Online]. Available: https://gdpr-info.eu/.

[35] ISO (International Organization for Standardization), ISO/IEC 27000:2018 Information Technology — Security Techniques — Information Security Management Systems — Overview and Vocabulary, Accessed: Oct. 11, 2021. [Online]. Available: https://standards.iso.org/ittf/PubliclyAvailableStandards/c073906_ISO_IEC_27000_2018_E.zip, 2018.
[36] CIS (Center for Internet Security), Center for Internet Security Controls, Version 7.1, Accessed: Oct. 11, 2021. [Online]. Available: https://learn.cisecurity.org/20-controls-download?_gl=1*2ttlk*_ga*MjA0MDEzNDk4LjE2ODQyNTE4MDI.*_ga_N70Z2MKMD7*MTY4NDI1NDcwMS4yLjEuMTY4NDI1NDcxMy40OC4wLjA.*_ga_ZQVR7NM9HJ*MTY4NDI1NDcwMS4yLjEuMTY4NDI1NDcxMy4wLjAuMA, 2019.
[37] A. Bello, A. Maurushat, Technical and behavioural training and awareness solutions for mitigating ransomware attacks, in: Advances in Intelligent Systems and Computing vol. 1226, AISC, 2020, pp. 164–176, https://doi.org/10.1007/978-3-030-51974-2_14.
[38] S. El-Sayegh, L. Romdhane, S. Manjikian, A critical review of 3D printing in construction: benefits, challenges, and risks, Arch. Civ. Mech. Eng. 20 (2) (Jun. 2020) 34, https://doi.org/10.1007/s43452-020-00038-w.
[39] D. Yao, B. García de Soto, A preliminary SWOT evaluation for the applications of ML to cyber risk analysis in the construction industry, IOP Conf. Ser. Mater. Sci. Eng. 1218 (1) (Jan. 2022) 012017, https://doi.org/10.1088/1757-899X/1218/1/012017.
[40] M.S. Sonkor, B. García de Soto, Is your construction site secure? A view from the cybersecurity perspective, in: Proceedings of the 38th International Symposium on Automation and Robotics in Construction (ISARC), ISARC, 2021, pp. 864–871, https://doi.org/10.22260/isarc2021/0117.
[41] G.D. Goh, S.L. Sing, W.Y. Yeong, A review on machine learning in 3D printing: applications, potential, and challenges, Artif. Intell. Rev. 54 (1) (Jan. 2021) 63–94, https://doi.org/10.1007/s10462-020-09876-9.
[42] G. Shemov, B. García de Soto, H. Alkhzaimi, Blockchain applied to the construction supply chain: a case study with threat model, Front. Eng. Manag. 7 (4) (Dec. 2020) 564–577, https://doi.org/10.1007/s42524-020-0129-x.
[43] A. Sheikh, V. Kamuni, A. Patil, S. Wagh, N. Singh, Cyber attack and fault identification of HVAC system in building management systems, in: 2019 9th International Conference on Power and Energy Systems (ICPES), IEEE, Dec. 2019, pp. 1–6, https://doi.org/10.1109/ICPES47639.2019.9105438.
[44] Z. Pan, S. Hariri, J. Pacheco, Context aware intrusion detection for building automation systems, Comput. Secur. 85 (Aug. 2019) 181–201, https://doi.org/10.1016/j.cose.2019.04.011.
[45] M.U.R. Mohamed Shibly, B. García de Soto, Threat modeling in construction: an example of a 3D concrete printing system, in: 37th International Symposium on Automation and Robotics in Construction, Oct. 2020, https://doi.org/10.22260/ISARC2020/0087.
[46] P. Mell, K. Scarfone, S. Romanosky, Common vulnerability scoring system, IEEE Secur. Priv. Mag. 4 (6) (Nov. 2006) 85–89, https://doi.org/10.1109/MSP.2006.145.
[47] B.R.K. Mantha, Y. Jung, B. García de Soto, Implementation of the common vulnerability scoring system to assess the cyber vulnerability in construction projects, in: Creative Construction e-Conference 2020, Budapest University of Technology and Economics, Budapest, Hungary, 2020, pp. 117–124, https://doi.org/10.3311/ccc2020-030.
[48] B.R.K. Mantha, B. García de Soto, Assessment of the cybersecurity vulnerability of construction networks, Eng. Constr. Archit. Manag. 28 (10) (Nov. 2021) 3078–3105, https://doi.org/10.1108/ECAM-06-2020-0400.
[49] B. Workshop, et al., BLOOM: A 176B-Parameter Open-Access Multilingual Language Model, arXiv, Jun. 27, 2023. Accessed: Mar. 02, 2024. [Online]. Available: http://arxiv.org/abs/2211.05100.
[50] L. Ouyang, et al., Training Language Models to Follow Instructions with Human Feedback, arXiv, Mar. 04, 2022. Accessed: Mar. 02, 2024. [Online]. Available: http://arxiv.org/abs/2203.02155.
[51] D. Yao, SFT Training Dataset, GitHub Repository, 2023. Accessed: Apr. 28, 2023. [Online]. Available: https://github.com/Daniel-Yao-Chengdu/NLP-project/blob/master/paper-A%20Cybersecurity-focused%20Large%20Language%20Model%20for%20the%20Construction%20Industry:%20A%20Case%20Study%20on%20Risk%20Identification/SFT%20training%20dataset.pt.
[52] D. Yao, Reward Model Training Dataset, GitHub Repository, 2023. Accessed: Apr. 28, 2023. [Online]. Available: https://github.com/Daniel-Yao-Chengdu/NLP-project/blob/master/paper-A%20Cybersecurity-focused%20Large%20Language%20Model%20for%20the%20Construction%20Industry:%20A%20Case%20Study%20on%20Risk%20Identification/Reward%20model%20training%20dataset.pt.
[53] D. Yao, RL Fine-tuning Dataset, GitHub Repository, 2023. Accessed: Apr. 28, 2023. [Online]. Available: https://github.com/Daniel-Yao-Chengdu/NLP-project/blob/master/paper-A%20Cybersecurity-focused%20Large%20Language%20Model%20for%20the%20Construction%20Industry:%20A%20Case%20Study%20on%20Risk%20Identification/RL%20fine-tuning%20dataset.pt.
[54] C. Raffel, et al., Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer, arXiv, Sep. 19, 2023. Accessed: Mar. 02, 2024. [Online]. Available: http://arxiv.org/abs/1910.10683.
[55] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning. Erscheinungsort nicht ermittelbar: Alanna Maldonado, 2023. ISBN: 978-973-23-4552-8.
[56] H. Zhang, et al., Fine-Tuning Pre-Trained Language Models for Few-Shot Intent Detection: Supervised Pre-Training and Isotropization, arXiv, May 26, 2022. Accessed: Mar. 02, 2024. [Online]. Available: http://arxiv.org/abs/2205.07208.
[57] Y. Yu, S. Zuo, H. Jiang, W. Ren, T. Zhao, C. Zhang, Fine-Tuning Pre-Trained Language Model with Weak Supervision: A Contrastive-Regularized Self-Training Approach, arXiv, Mar. 30, 2021. Accessed: Mar. 02, 2024. [Online]. Available: http://arxiv.org/abs/2010.07835.
[58] J. Shlens, Notes on Kullback-Leibler Divergence and Likelihood [object Object], 2014, https://doi.org/10.48550/ARXIV.1404.2000.
[59] B. Lowerre, R. Reddy, The harpy speech recognition system: performance with large vocabularies, J. Acoust. Soc. Am. 60 (S1) (Nov. 1976) S10–S11, https://doi.org/10.1121/1.2003089.
[60] R. Likert, A technique for the measurement of attitudes, Arch. Psychol. 22 (140) (1932) 55.
[61] D. Banerjee Chattapadhyay, J. Putta, R.M.P. Rao, Risk identification, assessments, and prediction for mega construction projects: a risk prediction paradigm based on cross analytical-machine learning model, Buildings 11 (4) (Apr. 2021) 172, https://doi.org/10.3390/buildings11040172.
[62] A. Gondia, A. Siam, W. El-Dakhakhni, A.H. Nassar, Machine learning algorithms for construction projects delay risk prediction, J. Constr. Eng. Manag. 146 (1) (Jan. 2020) 04019085, https://doi.org/10.1061/(ASCE)CO.1943-7862.0001736.
[63] N. Xia, R. Zhong, C. Wu, X. Wang, S. Wang, Assessment of stakeholder-related risks in construction projects: integrated analyses of risk attributes and stakeholder influences, J. Constr. Eng. Manag. 143 (8) (Aug. 2017) 04017030, https://doi.org/10.1061/(ASCE)CO.1943-7862.0001322.
[64] University of Defense, Military Aacademy, Belgrade, N. Kovačević, A. Stojiljković, University of Novi Sad, Faculty of Economics, Subotica, M. Kovač, and "Educons" University, Faculty for Project and Innovation Management, Belgrade, Application of the matrix approach in risk assessment, Oper. Res. Eng. Sci. Theory Appl. 2 (3) (Dec. 2019) 55–64, https://doi.org/10.31181/oresta1903055k.
[65] Accenture Security, 2020 Cyber Threatscape Report I, Accenture, Feb. 2020. Accessed: Mar. 01, 2024. [Online]. Available: https://www.accenture.com/content/dam/accenture/final/capabilities/technology/security/document/11177%20Cyber%20Threatscape%20Report_Digital_AW_SH.pdf.
[66] C. Cop, Phishing Email Detection, Accessed: Feb. 10, 2024. [Online]. Available: https://www.kaggle.com/datasets/subhajournal/phishingemails/data, 2023.
[67] T. Zhang, V. Kishore, F. Wu, K.Q. Weinberger, Y. Artzi, BERTScore: Evaluating Text Generation with BERT, arXiv, Feb. 24, 2020. Accessed: Mar. 02, 2024. [Online]. Available: http://arxiv.org/abs/1904.09675.
[68] D. Deutsch, D. Roth, Understanding the extent to which content quality metrics measure the information quality of summaries, in: Proceedings of the 25th Conference on Computational Natural Language Learning, Association for Computational Linguistics, 2021, pp. 300–309, https://doi.org/10.18653/v1/2021.conll-1.24. Online.
[69] B.R. Lashley, C.F. Bond, Significance testing for Round Robin data, Psychol. Methods 2 (3) (Sep. 1997) 278–291, https://doi.org/10.1037/1082-989X.2.3.278.
[70] M. Friedman, The use of ranks to avoid the assumption of normality implicit in the analysis of variance, J. Am. Stat. Assoc. 32 (200) (Dec. 1937) 675–701, https://doi.org/10.1080/01621459.1937.10503522.
[71] F. Wilcoxon, Individual comparisons by ranking methods, Biometrics 1 (6) (Dec. 1945) 80, https://doi.org/10.2307/3001968.
[72] C. Spearman, The proof and measurement of association between two things, Am. J. Psychol. 100 (3/4) (1987) 441, https://doi.org/10.2307/1422689.
[73] D. Yao, B. García de Soto, M. Wilkes, Identifying cyber risk factors associated with construction projects, Soc. Sci. Res. Netw. (2023), https://doi.org/10.2139/ssrn.4648243.