

Can ChatGPT Perform a Grounded Theory Approach to Do Risk Analysis? An Empirical Study

Yaxian Zhou, Yufei Yuan, Kai Huang & Xiangpei Hu

To cite this article: Yaxian Zhou, Yufei Yuan, Kai Huang & Xiangpei Hu (2024) Can ChatGPT Perform a Grounded Theory Approach to Do Risk Analysis? An Empirical Study, *Journal of Management Information Systems*, 41:4, 982-1015, DOI: 10.1080/07421222.2024.2415772

To link to this article: <https://doi.org/10.1080/07421222.2024.2415772>

View supplementary material 

Published online: 03 Jan 2025.



Submit your article to this journal



Article views: 1670



[View related articles](#)

View Crossmark data 

Citing articles: 1 View citing articles



Can ChatGPT Perform a Grounded Theory Approach to Do Risk Analysis? An Empirical Study

Yaxian Zhou^a, Yufei Yuan^{}^b, Kai Huang^b, and Xiangpei Hu^a

^aSchool of Economics and Management, Dalian University of Technology, Dalian, P. R. China; ^bDeGroote School of Business, McMaster University, Hamilton, Ontario, Canada

ABSTRACT

Grounded theory is a widely used scientific method for generating theories from qualitative data analysis. However, it is often time-consuming and requires professional training. Generative artificial intelligence, such as ChatGPT, excels in understanding and analyzing text, making it a valuable tool for qualitative research. This research proposes a novel approach to guide ChatGPT using the grounded theory method for qualitative data analysis and to design rigorous metrics for evaluating its performance. Using risk analysis as a case study, we compare ChatGPT's results with those obtained through manual methods. Our findings show that, with expert guidance, ChatGPT can effectively perform the grounded theory method, achieving results comparable to those of human analysts. To maximize its potential, researchers should properly guide ChatGPT in performing required tasks, rigorously evaluate its outputs, and ensure high-quality results. This approach can significantly enhance the efficiency and quality of qualitative data analysis.


KEYWORDS

generative AI; artificial intelligence; ChatGPT; grounded theory; risk analysis; qualitative analysis

Introduction

Grounded theory (GT) is an iterative and integrated process of data collection, analysis, coding, and conceptualization that culminates in theory generation [121]. Due to the capability of providing rich explanations of phenomena, discovering new theories, and the suitability across various fields, GT has become one of the most influential and widely used approaches in both of the social and natural sciences, including information systems (IS) research. Since the early 1990s, GT research has been quite prevalent in the mainstream journals of information systems (IS) for developing context-based, process-oriented descriptions and explanations of information systems phenomena [13, 78, 82, 83, 114, 115]. Given the call for theory building and development in IS, GT is increasingly being used in IS research [11, 17, 68]. However, GT relies on researchers' judgment and manual coding, leading to the data analysis process time-consuming and labor-intensive. Although the computational GT [84] proposed in 2020 incorporates computer-assisted text analysis techniques to mitigate the shortcomings of manual GT, it is still less advanced, because it requires specific machine learning domain knowledge (e.g., text mining

CONTACT Yufei Yuan  yuanfuyuf@mcmaster.ca  Michael G. DeGroote School of Business, McMaster University, 1280 Main Street West, Hamilton, ON L8S 4E8, Canada

 Supplemental data for this article can be accessed online at <https://doi.org/10.1080/07421222.2024.2415772>

© 2025 Taylor & Francis Group, LLC

knowledge) and sufficient training data to improve the accuracy of the analysis. Therefore, it is necessary to propose an approach that not only mitigates the limitations of manual GT, but also reduces the needs for training the text mining algorithm.

Chat Generative Pre-Trained Transformer (ChatGPT) is an advanced artificial intelligence (AI) large language model (LLM) released by a research company OpenAI [94] in November 2022. Having been pre-trained on vast corpus of data using deep learning and reinforcement learning algorithms, ChatGPT is capable of understanding the nuances of languages and generate highly accurate responses without any training and fine-tuning [97, 125]. Since its release, it has gained fastest-growing applications in various domains and tasks [16, 19, 36], and has exhibited human-level performance on many professional benchmarks [125]. The capability of generating human-level results in real time without training data and text mining knowledge as prerequisites makes ChatGPT a good complement to GT, which inspires us to investigate whether ChatGPT can perform GT to do risk analysis. We are interested in the extent to which ChatGPT can perform GT under the guide of humans. Therefore, we propose a ChatGPT-performed GT approach.

To demonstrate our approach, we take the risk analysis of electric power system blackouts as a case study. Risk analysis is fundamental to a variety of academic disciplines and is applicable to almost every aspect of life [100], such as health management [71], safety management [100], financial management [119] and information security management [118]. The purpose of risk analysis is to produce knowledge related to theories regarding how to understand, characterize, and govern risk [7]. The electric power system blackout is an important issue which has garnered wide attention from governments, industries, and academia. Blackout risk analysis aims to explore the causes and effects of blackouts and to provide decision support for preventing and responding to blackouts. The existing literature mainly employs statistical analysis, case study, and engineering models [75, 132] to analyze blackout events. Statistical charts are mainly for descriptive analysis of blackouts (e.g., summarizing the number of blackouts). Case study is more suitable for limited number of blackouts. Engineering methods focus on technical details of blackouts. GT is one of the most common approaches in the field of risk analysis, which has been mainly applied to identifying and classifying risk factors to develop risk terminology and taxonomy [83, 121, 136], exploring the relationships between risk factors [33, 127], and summarizing the measures of mitigating risks [31]. Due to the insufficient training data of blackouts, the emerging techniques such as machine learning and natural language processing (NLP) models are difficult to provide satisfactory results [58]. Taking the advantage of newly developed ChatGPT technology to mitigate these limitations, we propose a ChatGPT-performed GT approach to conduct the blackout risk analysis.

The ChatGPT-performed GT approach consists of five steps including information extraction, open coding, axial coding, selective coding, and data saturation test. We craft prompt patterns for each step to guide ChatGPT to perform corresponding tasks, and design quantitative metrics to evaluate ChatGPT's performance. In the empirical study of blackout risk analysis, we collect the text documents of 194 global blackout cases from multiple data sources, including investigation reports, technical reports, news reports, academic literature, and Wikipedia. We design and conduct comparative experiments in different scenarios, including different number of blackout cases processed by ChatGPT each time, and different number of examples provided for ChatGPT. To establish

a comprehensive understanding of ChatGPT's capabilities and limitations in performing GT, we discuss and summarize ChatGPT's performance in each step of our approach.

Compared with the existing literature, this research makes five contributions. First, this research proposes the novel idea of using ChatGPT to perform GT. To the best of our knowledge, this is the first research that integrates ChatGPT and GT. This research incorporates the advanced tool ChatGPT, which not only makes the data coding process of GT more efficient and reliable, but also can be used by humans without text mining knowledge. Second, this research designs a series of standardized and generalized prompt patterns to guide ChatGPT to perform GT, which can be applied to various domains without text mining knowledge. Third, this research proposes quantitative metrics to evaluate ChatGPT's performance, providing comprehensive perspectives to compare and understand the similarities and differences between ChatGPT and humans. Fourth, this research applies the ChatGPT-performed GT approach to blackout risk analysis based on text data, which provides a novel tool and perspective for exploring the risks of blackouts. The ChatGPT-performed GT approach can be flexibly adaptable to other risk analysis issues for risk management. Finally, this research introduces ChatGPT as a research assistant for human to use GT, which greatly promotes the efficiency and quality of GT research and as a result contributes to theory building and development in IS research. Researchers can apply our approach to theoretical conceptualization works as needed.

Related Work

This section reviews the existing literature related to this research, including GT, ChatGPT, and risk analysis.

Grounded Theory

GT was first formally introduced by two American sociologists Glaser and Strauss in 1967 for discovering theories through social research [48]. GT is designed to construct a theory that provides an abstract understanding of one or more core concerns in the studied world [24]. As one of the most popular and influential approaches, GT has spread from its original use in social science to other various disciplines and research fields [24, 79], such as health care [31], safety science [131], disaster [138] and accident management [33], business management [136], engineering construction [67], computer science [108], and information systems [68]. GT is summarized as an iterative and integrated process of data collection, analysis, coding, and conceptualization which culminates in theory generation [121]. In this process, data coding and conceptualization are the most important steps [46]. The theory of a phenomenon generated by GT mainly consists of concepts, categories, and their relationships [5, 29]. In contrast to many other approaches aiming at verifying the existing theories, GT focuses on generating theories from data. Therefore, GT can be applied in cases where there is no or limited existing theory that explains the phenomenon being studied [74]. In the IS field, GT has been frequently applied to exploring technological change and socio-technical behavior in emerging research domains [83]. GT has proved to be useful in developing theories, models and rich descriptions of IS phenomena, which are highly valued by the IS community [13, 78, 82, 83, 114, 115]. GT is considered to offer a compelling and appropriate research methodology for understanding IS-related

phenomena and developing IS-specific theory [11, 17, 68]. According to different development stages and philosophical backgrounds, there are three major variations of GT in the existing literature [4, 5].

Classic GT (or Glaserian GT) was introduced by Glaser and Strauss in 1967 for discovering theory from qualitative research of death and dying in health institutions [48]. Since then, Glaser has published a series of work to further elaborate on how to generate theories using GT [44, 45, 47]. Glaser and Strauss's groundwork, and Glaser's subsequent work are considered as the classic GT. Classic GT emphasizes that concepts and theories emerge directly from data, not rely on priori assumptions or frameworks [5, 48]. The data analysis processes of classic GT mainly consist of substantive coding, theoretical coding and continuous comparisons. For more details on classic GT, see Glaser's work [45, 47].

Systematic GT (or Straussian GT) was introduced by Strauss and Corbin [104] in 1990. Systematic GT provides an analytical framework for data coding to generate theories, which differs from Glaser's work that the theory emerges naturally from data. The data coding framework consists of open coding, axial coding, and selective coding [30], which extends the two-level coding process of classic GT to three levels. Systematic GT is more structured than classic GT [64], which provides an analytic data coding framework to guide researchers to conduct GT [24]. For more details on systematic GT, refer to existing literature [105, 106].

Constructivist GT (or Charmazian GT) was explicitly introduced by Charmaz [21] in 2000, which acknowledges the researcher's involvement in the construction and interpretation of data [22]. Constructivist GT insists that the theory is constructed through researchers' past and present experiences and research practices [22], which challenges the assumption of no a priori knowledge in classic GT. Furthermore, constructivist GT holds that the data coding framework of systematic GT is overly prescriptive and, therefore, provides an alternative data coding framework consisting of line-by-line coding and focused coding [24]. For more details on constructivist GT, refer to existing literature [23, 24].

Previous research [4, 5, 12, 57] provides detailed discussions and comparisons on the three major variations of GT. In addition, other new GT variants have evolved over time, such as situational GT [28], visual GT [63], transformational GT [95], and informed GT [111]. To mitigate the time-consuming and labor-intensive limitations of human coding in traditional GT, computational GT was proposed by Nelson [84] in 2020, which incorporates computer-assisted text analysis techniques into GT. Compared with human coding, computer-assisted techniques are more reliable and efficient. Nelson [84] combined three computer-assisted text analysis techniques into computational GT, including lexical-based techniques, text classification techniques, and natural language processing (NLP) techniques. The framework of computational GT consists of three steps: pattern detection, pattern refinement, and pattern confirmation. In these steps, the combination of computer techniques and human knowledge makes the theory-building and interpretation-rich tradition of GT more suitable for contemporary data-rich problems [84]. However, the research of computational GT [84, 85, 87] is in its infancy, and the text analysis techniques incorporated into computational GT are less advanced, which still requires specific knowledge in text mining or NLP techniques. The ChatGPT-performed GT approach proposed in this research improves computational GT, which incorporates a more advanced tool without specific text mining knowledge and data training as prerequisites.

ChatGPT

Generative AI refers to computational techniques capable of generating seemingly new, meaningful content such as text, images, or audio from training data [38]. ChatGPT is an advanced AI LLM released by a research company, OpenAI [94], in November 2022, which provides a browser-based interface to have access to the broad set of functionalities of generative AI models through simple queries [107]. ChatGPT is capable of engaging with users in a human-style conversations [2, 125], and remembering what users said earlier to realize continuous dialogue [125]. The inputs to ChatGPT by humans are referred to as “prompts” [88]. By designing effective prompts (i.e., prompt engineering techniques), humans can guide ChatGPT to provide more accurate and relevant responses [94]. Having been pre-trained on vast corpus of data using deep learning and reinforcement learning algorithms, ChatGPT is capable of understanding the nuances of languages and generate highly accurate responses even in complex and ambiguous contexts [97], and exhibits human-level performance on many professional benchmarks [125]. Therefore, ChatGPT is capable of handling a variety of tasks, such as extracting information, answering questions, writing essays and computer codes, editing manuscripts, translating languages, and so on [2, 35, 103, 125]. ChatGPT’s outstanding capabilities such as context understanding, task adaptability, and zero-shot learning revolutionize human-machine interactions [94] and have huge implications on academic research [35, 37]. LLMs represented by ChatGPT are substantially affecting and transforming social science research, which allows social scientists to break from traditional research methods and approach their work in innovative ways [52]. Since being released, ChatGPT has rapidly gained fastest-growing applications in various domains, such as biomedical and health [112], human resource management [16], finance research [36], account reconciliation [129], supply chains [39], and so on [9, 97, 120]. It is also reported recently that generative AI tools can summarize and extract the key points from research articles in succinct language and simulate abductive reasoning for research discovery [49].

ChatGPT is built on GPT models that have been continuously upgraded from GPT-1 to GPT-4o. ChatGPT is originally powered by GPT-3.5-Turbo model. With the release of GPT-4 model in March 2023, both GPT-3.5 and GPT-4 models are optimized for chat and available by employing prompt engineering techniques. GPT-4 is the large multimodal model of OpenAI with border general knowledge and advanced reasoning capabilities, which can resolve problems with higher accuracy than any previous GPT model [89]. Experiments have shown that GPT-4 outperforms the existing LLMs on most of the natural-language-processing tasks and exceeds most of the reported state-of-the-art systems [92]. Furthermore, GPT-4’s performance is strikingly close to human-level performance on various tests and academic benchmarks [15, 125]. For more details about ChatGPT, please refer to existing literature reviews [2, 9, 32, 43, 76, 94, 97, 102, 125].

To the best of our knowledge, ChatGPT has not been combined with GT in the existing literature. The five steps of our proposed approach are related to the following literature: the causal information extraction step is related to the literature on ChatGPT in information extraction and causal reasoning; the open coding, axial coding, and selective coding steps are related to the literature on ChatGPT in data coding; the saturation test step is related to the literature on ChatGPT in data annotation. The literature on information extraction tested ChatGPT on the tasks of entity recognition, relationship extraction, and event

extraction [25, 122, 135], revealing that ChatGPT can achieve high accurate results and provide high-quality explanations compared with humans [55, 70, 140]. The literature on causal reasoning applied ChatGPT to analyze causal relationships in customer reviews [91], disease diagnosis [113], and cyber security vulnerability [66], revealing that ChatGPT outperforms existing algorithms on a majority of causal tasks, like pairwise causal reasoning tasks [10, 62, 139]. The literature on data coding applied ChatGPT to capture the core content in qualitative data [81], such as producing main topics for a research project [81], and identifying common themes from statements [54], suggesting that ChatGPT might serve as an assistant in the traditional coding process [81]. The literature on data annotation applied ChatGPT to reproduce existing label annotations in the way of label selection, revealing that ChatGPT does have the potential to data annotation tasks [141]. Although the existing research has validated that ChatGPT does have the potential to handle these tasks related to GT, these tasks are not the same as those in this research. Taking the data coding step as an example, existing literature utilized simple prompts (e.g., “What is the main topic?”) to guide ChatGPT to extract core content, without designing systematic prompts based on qualitative analysis theory to execute the three coding steps. Our ChatGPT-performed GT approach can provide a demonstration of how to guide ChatGPT to perform qualitative analysis.

As an emerging AI model in the end of 2022, the application fields of ChatGPT remain to be further expanded. The capability of generating human-level results in real time without training data and specific text mining knowledge as prerequisites makes ChatGPT a good complement to GT and risk analysis. Together with the significant performance improvement of GPT-4 released in March 2023, these facts inspire us to apply ChatGPT to GT, which not only exploits the potential of ChatGPT but also mitigates the time-consuming and labor-intensive limitations of manual GT and text mining knowledge-dependent limitation of computational GT.

Risk Analysis

Risk analysis is a systematic process which focuses on understanding the nature of risk in a given situation and expressing the risk together with the underlying knowledge base [53]. The main purposes for conducting risk analysis include identifying the internal and external risks, rating the likelihood of the risks, rating the potential impact of the risks, and identifying actions that could help mitigate the risks [116]. As an important part of risk management, the process of risk analysis consists of identification of triggering events, cause analysis, effect analysis, and construction of risk pictures (see Figure 1) [6].

In the existing literature, GT is one of the important approaches for risk analysis. GT is applied in three aspects of risk analysis: 1) identifying and classifying risk factors to develop risk terminology and taxonomy; 2) exploring the relationships between risk factors; and 3) summarizing the measures to reduce risks. The representative research in the first aspect consists of identifying the risk factors of the infectious disease epidemic [40, 72], project construction [34, 67, 109, 137], software engineering [108], workers' unsafe behaviors [56, 127, 131], logistics transformation [73], supply chains [142], business innovation [61, 136], and various accidents and disasters [33, 133, 138]. The representative research in the second aspect consists of the risk transmission mechanism of cluster epidemic [71], the causations of subway operation [33], ship grounding accidents [56], and fake news [42]. The

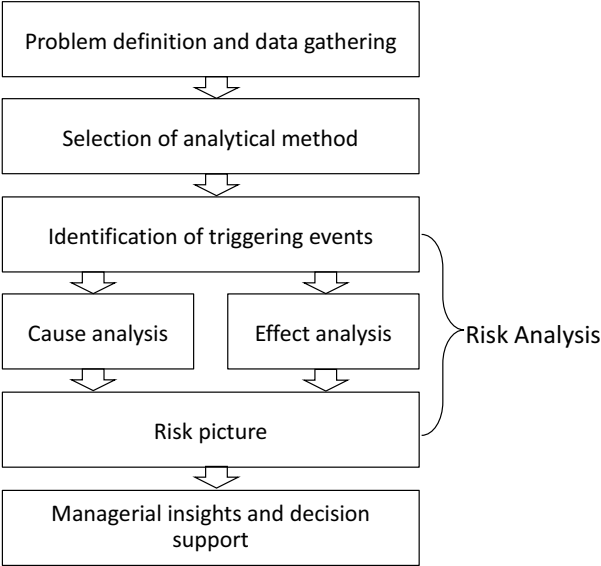


Figure 1. The main steps of the risk management process (modified after [6]).

representative research in the third aspect consists of summarizing the nursing strategies to prevent people from falling [31], the emergency capabilities for coal mine risks [126], and formulating flood risk management plans [14]. In summary, GT is mainly applied in the first aspect of risk identification and classification, which is implemented by manual coding of GT based on text documents.

With the advent of ChatGPT, some studies have validated the feasibility of applying ChatGPT to risk analysis in specific fields. In the health care domain, ChatGPT has been utilized to extract disease factors from collected papers, indicating its potential to construct disease risk databases [27]; ChatGPT has also been applied to assess suicide risk, suggesting its potential to estimate the likelihood of suicide attempts in a manner similar to assessments provided by professionals [69]. In the construction engineering domain, ChatGPT has demonstrated a superior ability to generate comprehensive risk management plans for construction projects [86]. ChatGPT has also been suggested for identifying potential risks in finance [43, 94] and supply chain [9], and for preventing and reducing natural disasters through its knowledge acquisition, interactive learning, and simulation capabilities [128]. However, research on the application of ChatGPT in risk analysis is still in its infancy, and some research suggestions have not yet been implemented. Moreover, there is no existing research using ChatGPT to perform GT approach to do risk analysis.

Risk analysis is interdisciplinary and significant for many areas, such as health risk [26], aviation risk [100], financial risk [119], project risk [77], and information security [118]. Blackout risk analysis is an important field, which aims to figure out the causes and effects of blackouts and provide decision-making support for preventing and responding to blackouts. Aligning with the main steps of risk analysis in Figure 1, the existing literature related to the blackout risk analysis can be subdivided into three aspects. The first is to identify the initiating causes of blackouts, which focuses on the classification of initiating causes. For example, some literature [3, 65, 134] classified the initiating causes of blackouts into macro

categories; some literature [101] focused on a certain macro category and further classified it into micro categories. The second is to describe the detailed process of blackouts [110]. The third is to analyze the impacts of blackouts, such as public health, economic and social impacts [8]. Most of the literature on this issue comes from the field of electrical engineering. Statistical analysis, case study, engineering models such as simulation models and topology models are the most common approaches [75, 132]. Statistical analysis especially the statistical chart provides descriptive analysis for the initiating causes of blackouts (e.g., the number of blackouts corresponding to different causes). Case study is more suitable for the limited number of blackouts. Engineering methods focus on the technical details of blackouts. These approaches have limitations in conducting blackout risk analysis based on text data. Furthermore, since the publicly available blackout text data cannot provide sufficient training data, the existing machine learning algorithms and NLP techniques are difficult to exhibit satisfactory performance [58]. Therefore, a more advanced approach is required, which is capable of effectively extracting blackout risk information from text data and reduces the needs for training machine learning algorithms.

In a broader way, many fields related to risk analysis require an approach that can generate theory from data and does not require training algorithms with insufficient risk data. On the one hand, risk analysis is a science producing knowledge related to concepts and theories regarding how to understand, characterize and govern risk [7]. On the other hand, risk events are infrequent and risk data usually involves confidential information, leading to limited accessible risk data. These two problems of risk analysis can be solved by the combination of GT and ChatGPT. GT is specifically designed for theory construction; pre-trained ChatGPT can achieve good results in small-scale risk data without training and fine-tuning [125]. Therefore, we propose the ChatGPT-performed GT approach, and employ the blackout risk analysis to demonstrate our approach. To the best of our knowledge, this is the first research that integrates ChatGPT, GT, and risk analysis.

Research Methodology

In this section, we propose the ChatGPT-performed GT approach, and craft prompt patterns for each step of the approach to guide ChatGPT performing corresponding tasks. Then, we design diverse quantitative metrics to evaluate ChatGPT's performance in each step of our approach.

ChatGPT-performed GT Approach

This research selects the systematic GT introduced by Strauss and Corbin [104] for ChatGPT, since it provides a structured data coding framework about how to conduct GT [24]. The structured framework of GT can guide us to design clear and specific prompts for ChatGPT, and more specific prompts can guide ChatGPT to generate more accurate responses. Based on the systematic GT, we propose the ChatGPT-performed GT approach shown in Figure 2.

In addition to data collection and preprocessing, our approach mainly consists of five steps: information extraction, open coding, axial coding, selective coding, and data saturation test. Aligning with the risk analysis process, the information extraction step corresponds to the risk information extraction, which outputs the causes and effects of risk

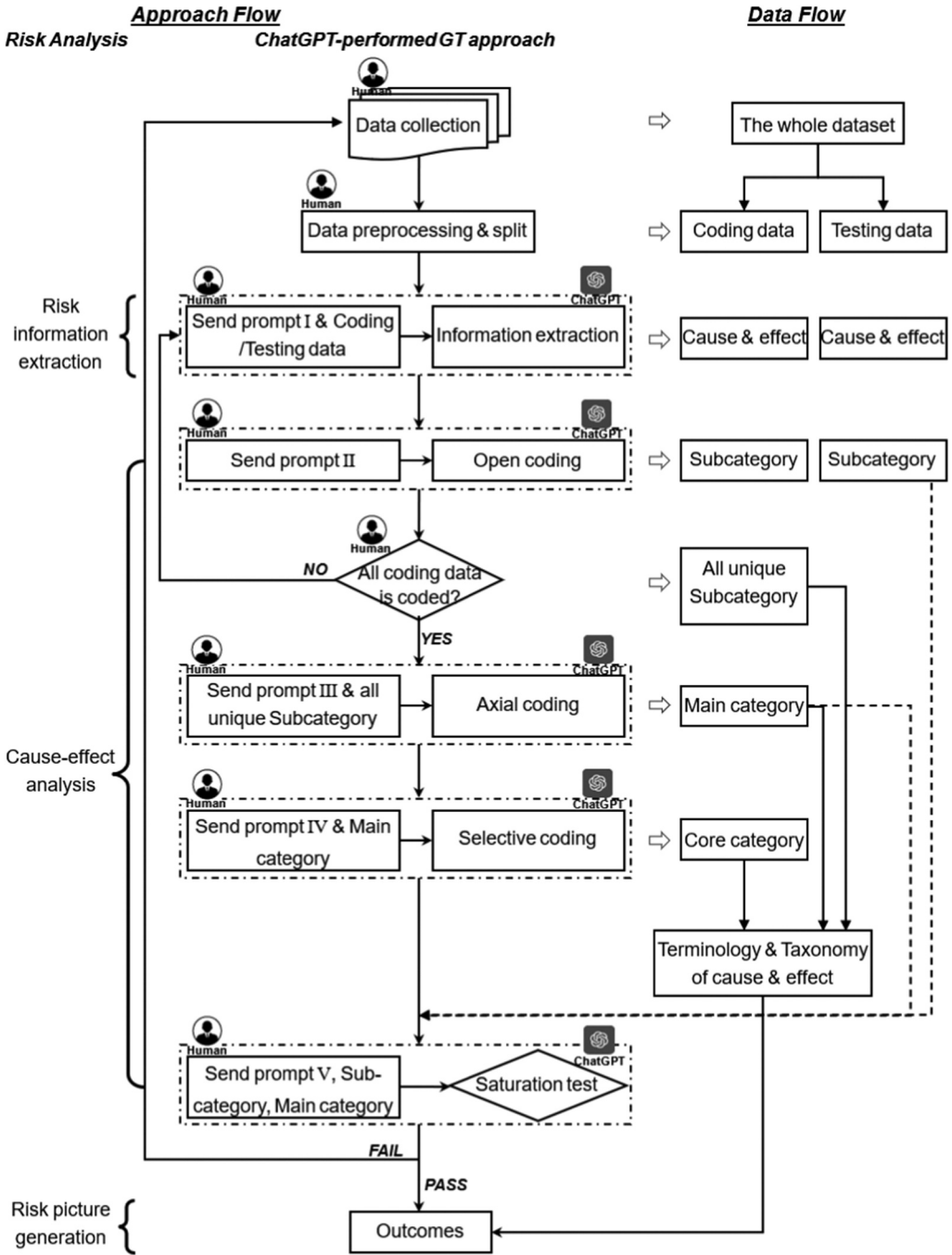


Figure 2. Framework of ChatGPT-performed grounded theory (GT) approach

events in the form of causal pairs; the data coding steps (i.e., open coding, axial coding, and selective coding) and data saturation test correspond to the cause-effect analysis, which encodes the causes and effects as subcategories, main categories, and core categories; finally, the terminology and taxonomy of causes and effects are output as a risk picture. Following

the guidelines of GT, data collection and data analysis are interrelated and iterative processes to avoid data redundancy caused by researchers spending a lot of time on collecting data [30], that is, data analysis (i.e., information extraction and three data coding steps) can be started after a small amount of data is collected. However, GT does not specify when each iteration of data collection starts and how long it lasts. Considering that the purpose of data collection in GT is to achieve data saturation, we employ a data saturation test as a measure for iterative data collection. Once data saturation is achieved, data collection and data analysis should cease. Instead, data collection and data analysis should start iteratively. In each step of our approach, humans interact with ChatGPT by sending prompts to it. A prompt is a set of task-specific instructions to guide ChatGPT to generate the desired output [123]. Prompt engineering is the means by which ChatGPT is programmed via prompts [123]. To learn more about prompt engineering, you may read the 2004 survey by Sahoo et al. [98]. Since prompt engineering can boost the performance of ChatGPT [94] and effectively communicate our intentions to ChatGPT, we employ prompt engineering to design prompt patterns for ChatGPT to perform GT for risk analysis. The prompt patterns consist of five main elements: role setting, task instruction, context description, input data, and output indicator [55]. The role setting assigns a role to ChatGPT to scope its responses to a designated point of view. The task instruction describes the specific task that ChatGPT should perform GT to complete. The context description introduces the background knowledge of risk analysis to ChatGPT for generating relevant responses. The input data refers to the data required by ChatGPT to perform tasks. The output indicator specifies the content and format (e.g., a paragraph or a table) of the desired output. The detailed steps and prompt patterns of our approach are discussed as follows.

Data Collection and Preprocessing

In this step, the dataset required for a research problem is collected and split into 80% for data coding (i.e., coding data), 20% for data saturation test (i.e., testing data).

Information Extraction

Information extraction refers to extracting the causes and effects of risk events from text data. Figure 3 shows our prompt pattern for information extraction. The sample response from ChatGPT is shown in Online Supplemental Appendix 2.

This prompt pattern starts with assigning the role of information extraction expert to ChatGPT. Then, we provide ChatGPT with the background knowledge of electric power system blackouts. The causes and effects can be represented by a long causal chain or several causal pairs split from a long causal chain. According to our preliminary trials, ChatGPT is more accurate at extracting causal pairs than longer causal chains. The existing research has also verified that splitting complex tasks into simpler subtasks can improve the accuracy of ChatGPT's response [88]. Therefore, we designate the task as extracting cause-effect pairs rather than causal chains. In order to distinguish causes and effects, we require ChatGPT to breakdown each cause-effect pair into a cause and an effect which are collectively referred as causal factors. The common causal factors from different cause-effect pairs should be named consistently. For convenient viewing, we require ChatGPT to provide the result in

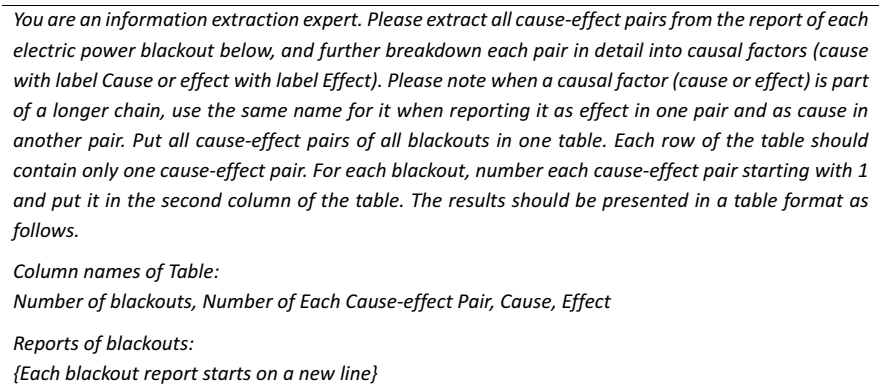


Figure 3. Prompt pattern I for information extraction.

the form of a table and explain the format of the table in detail. Finally, blackout reports are provided for ChatGPT as the input data for information extraction.

Data Coding

Data coding refers to employing codes (or concepts, categories) to conceptualize the extracted causes and effects of risk events to achieve maximum generality. Codes are simple word ideas that succinctly interpret phenomena reflected in the data [20]. The data coding of systematic GT consists of three steps: open coding, axial coding, and selective coding.

Open Coding

Open coding is a process to break down data and assign conceptualized codes to the data for capturing its meanings. In this research, the purpose of open coding is to refine the extracted causes and effects as concepts and summarize the concepts into subcategories. Before designing the prompt pattern, it is necessary to clarify whether the terminology “open coding” can be directly applied to describe the task for ChatGPT. Therefore, we required ChatGPT to explain open coding in the systematic GT approach and obtained an accurate response from ChatGPT, suggesting that ChatGPT is capable of understanding the open coding task and we can directly apply the terminology “open coding” without explanations in the prompt pattern.

As shown in Figure 4, the prompt pattern for open coding starts with assigning the role of GT expert to ChatGPT. Then, we require ChatGPT to conduct the open coding process, and insert the concepts and subcategories into the table which has been constructed in the previous step (i.e., information extraction). To avoid redundant codes, we remind ChatGPT to assign the same code to identical causal factors. In this prompt pattern, it is not necessary to provide the input data (i.e., the extracted causal factors), because the steps of open coding and information extraction are conducted in the same conversation with ChatGPT, and the results of information extraction can be remembered by ChatGPT as the input data for open coding. It should be noted that the ChatGPT service we are using has limitations in the maximum length of input data. If

You are an expert specializing in grounded theory. Please conduct the open coding process for every causal factor (cause and effect) in the third and fourth column of the above table, insert the specific concept for every cause into the fifth column (named CauseConcept) of the table, insert the specific concept for every effect into the sixth column (named EffectConcept) of the table, insert the broader subcategory for every cause into the seventh column (named CauseSubCategory) of the table, and insert the broader subcategory for every effect into the eighth column (named EffectSubCategory) of the table. If two or more causal factors are same or similar, please provide a same concept and subcategory for them. The results should be presented in a table format as follows.

Column names of Table:

Number of blackouts, Number of cause-effect Pairs, Cause, Effect, CauseConcept, EffectConcept, CauseSubCategory, EffectSubCategory

Figure 4. Prompt pattern II for open coding.

there are a large number of risk events, they need to be split and open coded in batches. After the open coding of all data is performed by ChatGPT, we collect all subcategories and remove duplicates for subsequent axial coding.

Axial Coding

Axial coding is a process to abstract the conceptualized results (i.e., subcategories) from open coding, and to cluster the associated subcategories into a more abstracted and generalized category named main category. As we did in the open coding step, we first determined ChatGPT's capability of understanding the terminology "axial coding," suggesting that this terminology can be directly applied in the prompt pattern.

As shown in Figure 5, the prompt pattern for axial coding starts with assigning the role of GT expert to ChatGPT. Then, we require ChatGPT to conduct the axial coding process. The input data refers to all unique subcategories obtained from the open coding step. The output results consist of main categories, the interpretation of each main category, and the subcategories clustered into each main category, which help us understand main categories and establish the relationships between main categories and subcategories.

You are an expert specializing in grounded theory. The subcategories related to the causal factors of blackouts are provided for you. Please conduct the axial coding process for the subcategories. List each main category for abstracting subcategories and explain it. List all subcategories clustered in this main category. The results should be presented in the following format.

Main category: explanation of the main category

{Subcategory 1, Subcategory 2, ...}

[Input data]

Subcategories:

{text}

Figure 5. Prompt pattern III for axial coding.

Selective Coding

Selective coding is a process by which all categories generated in the previous steps are unified around a core category [30]. The core category represents the central phenomenon of the research [30]. A conceptual model can be established to assist in integrating all categories and revealing their relationships. As we did in the open and axial coding steps, we determined ChatGPT’s capability of understanding the terminology “selective coding,” indicating that we can directly apply this terminology to assigning tasks for ChatGPT.

As shown in Figure 6, ChatGPT conducts selective coding directly without being provided a role and input data. This is because selective coding and axial coding are conducted successively in the same conversation with ChatGPT, and the results of axial coding (i.e., main categories and their definitions) can be remembered by ChatGPT as the input data of selective coding. The output results consist of a core category and its explanation, and descriptive analysis of the relationships between all the main categories and the core category. Finally, the three types of categories (i.e., subcategories, main categories, and core categories), and their relationships are referred to construct the terminology and taxonomy of the causes and effects of risk events (i.e., risk picture), which represents the result of ChatGPT-performed approach for risk analysis.

Data Saturation Test

The purpose of data saturation test is to densify the concepts and categories emerging in GT [12]. The test idea is to generate categories from the testing data and determine whether these categories are new compared to those obtained from the coding data. If the two set of categories are similar, the research successfully passes the saturation test, indicating that the steps of data collection and data coding should cease; If new categories emerge from the testing data, the research fails the test, indicating that the steps of data collection and data coding should continue until passing the test.

In this research, the remaining testing data is used to conduct the saturation test. First, ChatGPT performs the prompt patterns of information extraction and open coding on testing data to generate test subcategories. Second, to determine whether new subcategories emerge from these test subcategories, we employ all the existing main categories obtained from coding data to design prompt patterns. Our design idea is to require ChatGPT to determine whether each test subcategory can be labeled by the existing main categories. If all test subcategories can be labeled, it suggests no new subcategories emerging. If any test subcategory cannot be labeled, it suggests new subcategories emerging, and ChatGPT is required to generate a new main category for the new subcategory.

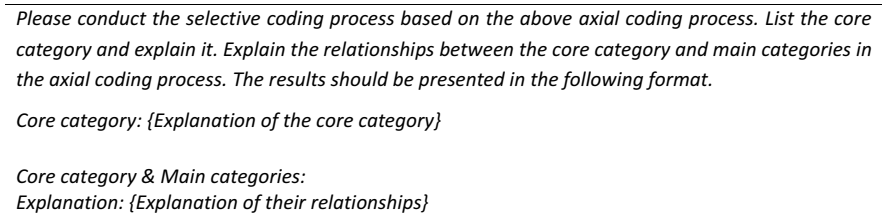


Figure 6. Prompt pattern IV for selective coding.

You are an expert specializing in grounded theory. The subcategories related to the causal factors of blackouts are provided for you. Please determine whether each subcategory can be labeled by the following one of main categories. If there are no appropriate main categories, please output "No appropriate label" and generate a better main category. The results should be presented in a table format as follows.

Column names of Table:

Subcategory, Label

[Input data]

Main categories:

[main category 1, main category 2, ... or main category n].

Subcategories:

{text}

Figure 7. Prompt pattern V for data saturation test.

As shown in [Figure 7](#), after providing the role of GT expert and the background knowledge of subcategories, the task of saturation test is assigned to ChatGPT based on the design idea. The input data consists of test subcategories generated in this step and main categories in the axial coding step. The output result is presented in the form of a table. If the phrase "No appropriate label" does not appear in the table, this research passes the saturation test. If the phrase "No appropriate label" appears in the table, this research may fail the saturation test and ChatGPT will recommend a new main category for the corresponding new subcategory. We can evaluate the reasonability of ChatGPT's results and obtain the final conclusion.

Evaluation Metrics

In this section, we propose three quantitative metrics: Consistency, Precision, and Recall, to evaluate the performance of ChatGPT in each step of the ChatGPT-performed GT approach. For all evaluation metrics, the results provided by domain experts (hereinafter called humans) are regarded as the benchmark for comparison. When performing the GT approach, human(s) should be given identical experimental datasets and prompts with ChatGPT. Humans provide the benchmark by individual or collaborative work. Compared with individual work, GT advocates collaborative work by humans who may share interest and/or experience in the same field [30], regardless of the number of collaborators. Therefore, humans can flexibly devise methods of collaboration as needed to achieve the benchmark for evaluating the performance of ChatGPT. Consistency is proposed to evaluate the degree of agreement between ChatGPT and humans. Precision [130] is employed to evaluate the proportion of codes (i.e., concepts or categories) obtained by ChatGPT that are also identified by humans. Recall [130] is employed to evaluate the proportion of codes (i.e., concepts or categories) obtained by humans that are also identified by ChatGPT. Consistency, Precision, and Recall are defined as follows.

$$\text{Consistency} = NC_{GPT \text{ and } MAN} / NC_{GPT \text{ or } MAN} \times 100\% \quad (1)$$

$$Precision = NC_{GPT \text{ and } MAN} / NC_{GPT} \times 100\% \quad (2)$$

$$Recall = NC_{GPT \text{ and } MAN} / NC_{MAN} \times 100\% \quad (3)$$

In the information extraction step, $NC_{GPT \text{ and } MAN}$ represents the number of causal pairs extracted by both ChatGPT and humans; $NC_{GPT \text{ or } MAN}$ represents the number of causal pairs extracted by ChatGPT or humans; NC_{GPT} represents the number of causal pairs extracted by ChatGPT; NC_{MAN} represents the number of causal pairs extracted by humans. Consistency represents the degree of agreement on causal pairs between ChatGPT and humans. Consistency = 100% indicates the full agreement of causal pairs extracted by ChatGPT and humans; Consistency < 100% indicates the partial agreement of causal pairs extracted by ChatGPT and humans. Precision represents the proportion of causal pairs extracted by ChatGPT that are identified by humans as “true” causal pairs. Recall represents the proportion of “true” causal pairs identified by humans that are also extracted by ChatGPT as causal pairs. The values of Precision and Recall range from 0 to 100%. The higher values of Precision and Recall indicate the more comparable capability of ChatGPT to humans in extracting “true” causal pairs.

In the open coding or axial coding step, $NC_{GPT \text{ and } MAN}$ represents the number of categories generated by both ChatGPT and humans, that is the number of human-generated categories whose meanings are captured by ChatGPT-generated categories; $NC_{GPT \text{ or } MAN}$ represents the number of categories generated by ChatGPT or humans; NC_{GPT} represents the number of categories generated by ChatGPT; NC_{MAN} represents the number of categories generated by humans. Consistency represents the degree of agreement on categories between ChatGPT and humans. Consistency = 100% indicates the full agreement of categories generated by ChatGPT and humans; Consistency < 100% indicates the partial agreement of categories generated by ChatGPT and humans. Precision represents the proportion of categories generated by ChatGPT that are also generated by humans. Recall represents the proportion of categories generated by humans that are also generated by ChatGPT. The higher values of Precision and Recall indicate the meanings of more human-generated codes can be captured by ChatGPT, suggesting the more comparable capability of ChatGPT to humans in abstracting categories. Since there is usually one core category in the selective coding step, we can directly evaluate this core category without employing these three metrics.

Experiments and Results Analysis

This section provides an overview of the experimental dataset and the experimental setup, followed by a detailed analysis and discussion of the experimental results.

Experimental Dataset

The data used by GT can be first-hand data through interviews or field notes, and texts and information from other sources such as historical documents, government records, or organizational information compiled for private discussion or public dissemination [22]. Following this guideline of GT, we collect the text data of 194 worldwide blackout cases occurred from 1965 to 2021 from official investigation reports, news reports, academic

literature, and Wikipedia. Since the raw text data contains a lot of content irrelevant to risk analysis, we conduct a rough reading of the raw data and screen out the content related to the blackout risk analysis as the research dataset. Our selection of the research data is for the demonstration purpose. As the founder of GT pointed out, the accuracy, trustworthiness or objectivity of research data is an inherent issue in qualitative research including GT [47]. However, GT does not provide standardized procedures for data collection and processing. It should be noted that for further research, a standardized or blind screening method will be needed to enhance the objectivity of research data. In the future, data gathering, screening, and verification may also be supported by advanced generative AI that integrated with searching engine such as Gemini in Google Cloud Vertex AI [51]. Among the research dataset, the text data of 150 blackout cases is selected as coding data for data coding analysis, while the remaining data of 44 blackout cases is employed as testing data for data saturation test.

Experimental Setup

Since the GPT-4 model is claimed to be the most advanced model and superior to the GPT-3.5 model [90], this research selects ChatGPT-4 to conduct the experiments. Three domain experts are invited to perform the systematic GT approach on the same experimental dataset and prompts. All domain experts are familiar with ChatGPT and GT approach. One of the domain experts performs the manual systematic GT approach to obtain the first version of results, and then these three domain experts discuss and update the first version of results and reach an agreement to provide a final result as the benchmark for evaluating ChatGPT's performance. Two aspects of comparative experiments are designed to evaluate ChatGPT's performance: the number of cases processed by ChatGPT each time, and the number of examples provided to ChatGPT.

The first comparative experiment is to compare ChatGPT's performance in processing different numbers of blackout cases (batch) each time. The number of blackout cases provided for ChatGPT each time is proportional to the length of the message. The existing research has shown that ChatGPT's responses can be affected by the length of the message it receives [91], which inspires us to compare ChatGPT's performance in one-blackout and few-blackout scenarios. A one-blackout scenario means that ChatGPT analyzes one blackout report each time. A few-blackout scenario means that ChatGPT analyzes several blackout reports (e.g., 3 blackouts, 5 blackouts) each time. Therefore, this research compares the performance of ChatGPT in 1-, 5-, 10- and 15-blackout scenarios. Taking the 5-blackout scenario as an example, for 150 blackout reports in coding analysis, ChatGPT needs to run 30 times to process all of these reports.

The second comparative experiment is to explore the impact of adding examples as task demonstrations (demo) on the performance of ChatGPT. This research compares ChatGPT's performance in the zero-shot and one-shot scenarios. A zero-shot scenario means that we do not provide any examples in the prompt pattern to guide ChatGPT on specific tasks. A one-shot scenario means that we provide one example in the prompt pattern to allow ChatGPT to learn and complete tasks by imitation [125]. ChatGPT has been pre-trained on vast amounts of data and can be generalized to the new data that it has not been specifically trained on. Therefore, ChatGPT is capable of conducting the zero-shot prompt without any training and fine-tuning [55]. When the zero-shot prompt does not

work on complex tasks, the one-shot prompt is recommended to guide ChatGPT to better performance. The one-shot prompt patterns for information extraction, open coding and axial coding are provided in [Figures 1.1-1.3](#) of Online Supplemental Appendix 1. Compared with the patterns in [Figures 3-5](#), these figures have one additional example (highlighted in bold), which is provided by domain experts. Each one-shot experiment employs the same example for consistency.

Results Analysis

In this section, we present and analyze ChatGPT’s performance in each step of ChatGPT-performed GT approach for risk analysis, according to the outcomes of all comparative experiments (i.e., 4 (batch) x 2 (demo) = 8 scenarios in total).

Performance on Information Extraction

The performance of ChatGPT in the information extraction step is evaluated by the values of Consistency, Precision, and Recall ([Table 1](#)). The highest value of Consistency (99.36%) suggests that ChatGPT and humans achieve 99.36% agreement on the causal pair extraction. The highest values of Precision (100%) suggest that all of the ChatGPT-extracted causal pairs are identified as “true” causal pairs by humans. The highest values of Recall (100%) suggest that all the human-identified causal pairs are extracted by ChatGPT. These three highest values indicate that ChatGPT demonstrates human-comparable capability in the information extraction step. For information extraction, when only one case is processed by ChatGPT each time, there are a minimal number of differences between zero-shot scenario and one-shot scenario, with Consistency (99.18% vs. 99.36%), Precision (99.18% vs. 99.36%), and Recall (100% vs. 100%), suggesting that ChatGPT’s performance is not affected by providing examples. However, as the number of blackout cases processed each time increases, a trend of performance decrease is observed in the values of Consistency and Recall, suggesting that the increasing number of blackout cases processed each time (batch workload) can lead to the decline of ChatGPT’s accuracy during the causal pair identification in information extraction.

Although the Precision value for information extraction is higher than 99%, there are five cases in which the ChatGPT-extracted causal pairs are false in some scenarios. [Table 2](#) summarizes all of the five false causal pairs identified by ChatGPT: the first

Table 1. The values (%) of Consistency, Precision and Recall in different scenarios and steps.

Shots		Zero shot				One shot			
Cases		1	5	10	15	1	5	10	15
Information Extraction	Consistency	99.18	98.51	93.21	89.64	99.36	96.88	96.10	95.00
	Precision	99.18	99.62	100.00	100.00	99.36	99.60	100.00	100.00
	Recall	100.00	98.88	93.21	89.64	100.00	97.25	96.10	95.00
Open Coding	Consistency	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
	Precision	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
	Recall	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
Axial Coding	Consistency	80.00	90.00	90.00	80.00	90.00	80.00	90.00	90.00
	Precision	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
	Recall	80.00	90.00	90.00	80.00	90.00	80.00	90.00	90.00

Table 2. Five false causal pairs identified by ChatGPT.

Sentences from blackout reports	Causes	Effects	Scenarios
A fire at a substation of the power plant triggered a 60-hour-long island-wide blackout beginning May 18.	60-hour-long island-wide blackout	Beginning May 18	zero shot-1 task
The blackout was caused by technical malfunction in the transmission line coming from Dam.	A transmission line coming from Dam	Technical malfunction	zero shot-1 task
An electrical blackout has been newly blamed on computer hackers, but was actually the result of a utility company's negligent maintenance of high voltage insulators on two transmission lines.	Computer hackers	Electrical blackout	zero shot-1 task one shot-5 task
The interaction of coincident threats - strong winds reaching hurricane force, temperatures about 0°C, heavy snow load on the lines and rain - gave rise to torn power lines and over 80 ruptured pylons.	Rain	Heavy snow load on the lines	zero shot-5 task one shot-1 task
This led to a blackout affecting 250,000 people who had to cope with electricity supply interruptions for nearly five days.	Blackout affecting 250,000 people	Electricity supply interruptions for nearly five days	one shot-1 task

Table 3. Comparison of causal pairs extracted by ChatGPT and Human.

Sentences from a report	Extracted by	Causes	Effects
The blackouts were caused by a sudden increase in electricity demand for air conditioning during a strong heat wave and a simultaneous explosion at the thermal power plant's substation.	Human	Heat wave	Electricity demand increase
		Electricity demand increase	Blackout
	ChatGPT	Explosion at a substation	Blackout
		Electricity demand increase during a heat wave	Blackout
		Explosion at a substation	Blackout

pair reflects the relationship between a blackout event and its occurrence time; the second pair reflects the fault and the location of this fault in the power system; the third pair is not the actual cause of the blackout according to the context; the fourth pair reflects the parallel relationship between the two causes of a blackout; the fifth pair reflects the parallel relationship between the two consequences of a blackout. It is worth noting that ChatGPT does not always make these mistakes. Taking the third pair as an example, ChatGPT can automatically provide the cause “Computer hackers” with an explanation “misattributed cause” in certain scenarios, suggesting that ChatGPT is capable of understanding the context and extracting the actual cause of this blackout case.

In addition to the two scenarios with a perfect Recall value (100%), there are six other scenarios with a Recall value between 89.64% and 100%, which means some of the human-identified causal pairs cannot be extracted by ChatGPT in these scenarios. This is because ChatGPT sometimes does not fully extract nested causal pairs. For example (Table 3), ChatGPT does not further extract “heat wave” from “electricity demand increases during a heat wave” as a cause of the blackout, while humans further break down “electricity demand increases during a heat wave” and extract “heat wave → electricity demand increase” as a causal pair.

Performance on Open Coding

The performance of ChatGPT in the open coding step is evaluated by the values of Consistency, Precision, and Recall (Table 1). The perfect Consistency values (100%) in all scenarios suggest that ChatGPT and humans achieve full agreement on subcategories. The perfect Precision values (100%) in all scenarios suggest that the meanings of all ChatGPT-generated subcategories are reflected in human-generated subcategories. The perfect Recall values (100%) in all scenarios suggest that the meanings of all human-generated subcategories are successfully captured by ChatGPT. The values of these three metrics indicate that ChatGPT demonstrates the human-comparable capability in the open coding step.

Compared with human-generated subcategories, ChatGPT-generated subcategories can be divided into three groups: 1) Similar subcategories which are the same as human-generated subcategories; 2) Finer subcategories which are more detailed than human-generated subcategories; and 3) Broader subcategories which are more extensive than human-generated subcategories. The human- and ChatGPT-generated subcategories are selected from different scenarios, including similar, finer, and broader subcategories (Table 4). As shown in Table 4, except for the human-generated subcategory “internal human error” (in bold), ChatGPT is capable of generating similar subcategories for all of the other human-generated subcategories, suggesting a high similarity between ChatGPT- and human-generated subcategories.

Performance on Axial Coding

The performance of ChatGPT in the axial coding step is evaluated by the values of Consistency, Precision, and Recall (Table 1). The highest Consistency value (90%) suggests that ChatGPT and humans achieve 90% agreement on the main categories.

Table 4. Comparison of subcategories generated by ChatGPT and Human.

Human-generated subcategories	ChatGPT-generated subcategories		
	Finer subcategories	Similar subcategories	Broader subcategories
Earthquake	—	Earthquake/Seismic activity	Natural events
Mass movement	Landslide collapse	Terrain disruption	Natural disasters
Storm	Winter storm	Storm	Environmental factors
Extreme temperature	Sudden temperature drop	Extreme temperature	Weather events
Flood	—	Flood	Weather events
Drought	—	Drought	Climatic events
Fire/explosion	Bushfire	Explosion/Fire	—
Tree interference	—	Tree interference	Environmental impact
Animal interference	Bird electrocution	Animal interference	—
Space weather	Geomagnetic disturbance	Space weather	Natural phenomenon
Generation fuel shortage	Fuel quality	Fuel shortage	Supply chain issue
Generation equipment shortage	Equipment aging	Resource shortage	Equipment failure
Demand overload	—	Demand surge	Grid stability issue
Internal human error	Operational mistake	—	Human error
External human error	Sabotage	External human error	Human error
Physical attacks	Electromagnetic attack	Physical damage	External attack
Cyber attacks	—	Cyber attack/Hacking	Security breach
Equipment defect/failure	Cable fault	Equipment defect/failure	—
Equipment fire/explosion	Substation fire	Equipment fire/explosion	Equipment failure
Support system failure	—	Control system failure	—
Electrical quality issue	Frequency drop	Power quality	—

Table 5. Comparison of some selected main categories generated by ChatGPT and Human.

Human-generated main categories	ChatGPT-generated main categories		
	Finer main category	Similar main category	Broader main category
Geophysical disasters, Meteorological disasters . . .	Weather-related factors	Meteorological factors	Natural and environmental disruptions
Technical issues	Electrical and equipment Failures	Technical issues	—
Human-unintentional errors, Human-intentional acts	Management and procedural issues	—	Human-caused disruptions
Supply issues, Demand issues	Resource and supply disruptions	Power demand and supply factors	Systemic and planning issues

The perfect Precision values (100%) in all scenarios suggest that the meanings of all ChatGPT-generated main categories are reflected in human-generated main categories. The highest Recall value (90%) suggests that the meanings of 90% of the human-generated main categories are successfully captured by ChatGPT. The values of these three metrics indicate that ChatGPT demonstrates an excellent capability in the axial coding step.

As the classification of subcategories in the open coding step, the main categories are also divided into three groups: similar, finer, and broader main categories. The human- and ChatGPT-generated main categories are selected from different scenarios, including similar, finer, and broader main categories (Table 5). As shown in Table 5, for the human-generated main categories “human-unintentional errors” and “human-intentional acts,” ChatGPT tends to generate one broader category “human-caused disruptions” to describe both of them; for some specialized human-generated main categories, such as “geophysical disasters,” “demand issues,” and “supply issues,” ChatGPT tends to generate border main categories such as “natural and environmental disruption” and “systematic and planning issues.” This suggests that ChatGPT is capable of capturing the meanings of human-generated main categories, but tends to generate broader main categories with lower similarity compared with humans.

Performance on Selective Coding

Based on the prompt pattern of this step (Figure 6), ChatGPT will generate a core category and its explanation, and the relationships between the core category and main categories. Since only one core category is generated in one scenario, the eight core categories in all scenarios are summarized in Table 6. According to our evaluation, although the eight core categories are different from each other, all of them correctly reflect the topic related to the experimental dataset. Notably, one core category underlined in Table 6 does not explicitly mention the blackout or the power system. The explanation of the core category generated by ChatGPT can further support the evaluation: The core category of “infrastructure vulnerabilities and external influences” serves to unify the various causes of blackouts, emphasizing the multifaceted vulnerabilities in the power infrastructure and the constant challenge of maintaining reliable power in the face of both inherent and external challenges. According to the explanation, it can be evaluated that ChatGPT is capable of correctly grasping the topic of the experimental dataset and generating diverse core categories, which provides alternative perspectives of the data.

Table 6. Core categories generated by ChatGPT.

	The number of blackouts processed each time			
	1	5	10	15
Zero-shot	Causes of blackout disruptions	Power system resilience	Power system disruptions	Blackout precursors
One-shot	Power system stability and reliability	System vulnerabilities in power infrastructure	Power system stability and reliability	Infrastructure vulnerabilities and external influences

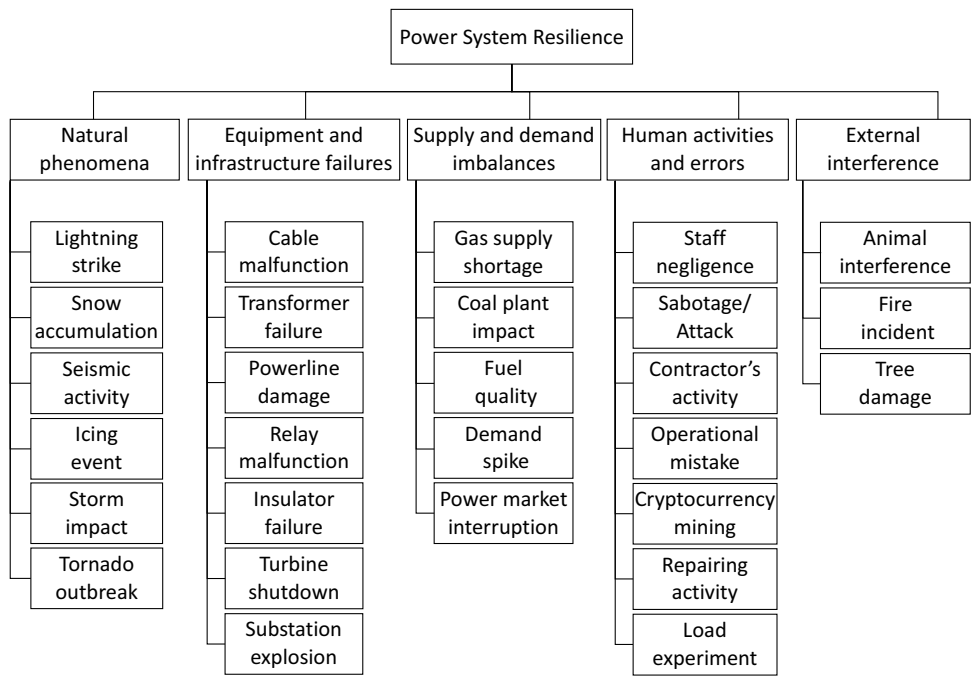


Figure 8. Emerging categories and their relationships generated by ChatGPT.

Figure 8 illustrates the terminology and taxonomy of the causes and effects of blackouts as the result of the GT approach, which is based on the three types of categories (subcategory, main category, and core category) and their relationships generated by ChatGPT. Due to the space limitation, only partial categories are selected to present. From the first to the third layer in Figure 8, there are 1 core category, 5 main categories, and 28 subcategories. The explanations for their definitions and relationships are provided by ChatGPT. For example, the main category “human activities and errors” is defined as human negligence and intentional acts that disrupt power systems, which includes 22 subcategories; the relationship between this category and the core category “power system resilience” is explained as follows: human errors can pose significant threats to the power grid; the resilience of the power system in this context refers to the ability to predict human-induced risks, protect against them and rapidly rectify issues. These explanations are evaluated by domain experts as reasonable and correct, suggesting that ChatGPT demonstrates the human-comparable capability in the selective coding step.

Table 7. The number of “No appropriate label” and subcategories for the 44 blackouts.

	The number of blackouts processed each time			
	1	5	10	15
Zero-shot	0/15	0/71	0/65	1/58
One-shot	0/74	1/57	0/54	0/40



Here's a table mapping each subcategory to its most appropriate main category:

Subcategory	Label
Asymmetric Condition	Equipment & Infrastructure Issues
Automatic Plant Shutdown	Equipment & Infrastructure Issues
Bird Electrocution	Environmental Factors
Brush Fire	Environmental Factors
Busbar Fault	Equipment & Infrastructure Issues
Cable Fault	Equipment & Infrastructure Issues
Cascade Tripping	Equipment & Infrastructure Issues
Circuit Breaker Issue	Equipment & Infrastructure Issues
Coastal Winds	Environmental Factors
Cold Temperature Demand	Environmental Factors
Communication Infrastructure Damage	Equipment & Infrastructure Issues
Crew Mistake	Human Factors & Interference
Cyclonic Storm	Environmental Factors
Earthquake	Environmental Factors
Explosion	No appropriate label (Potential label: Accidents & Unplanned Events)

Figure 9. The response of ChatGPT for data saturation test.

Performance on Data Saturation Test

Table 7 and Figure 9 present the results of data saturation test generated by ChatGPT. The task of this step is to determine whether new subcategories generate from the text data of 44 blackouts remained in the dataset. The subcategories of the 44 blackouts have been generated through the steps of information extraction and open coding. By comparing these subcategories with the existing main categories from the 150 blackouts, domain experts conclude that all subcategories can be labeled by the existing main categories and

no new subcategories emerge from the 44 blackouts, which indicates that this research passes the data saturation test. According to the prompt pattern designed for ChatGPT (Figure 7), if the phrase “No appropriate label” and an alternative main category label are generated in the response (Figure 9), it implies a new subcategory emerging and inconsistent conclusion with humans. Therefore, the number of “No appropriate label” and subcategories generated by ChatGPT are summarized in Table 7. Taking the result “0/15” as an example, 0 represents the number of “No appropriate label,” and 15 represents the number of subcategories correctly labeled by the existing main categories.

As shown in Table 7, the number of No appropriate label in two scenarios is 1 (in bold), suggesting that a new subcategory may emerge in these two scenarios. ChatGPT’s response in one of the scenarios (zero shot-15 task) is presented in Figure 9 for further evaluation. The subcategory “explosion” is labeled with “No appropriate label” followed with a new potential main category “Accidents & Unplanned events.” Furthermore, without human involvement, ChatGPT automatically provides the following detailed explanation: the tricky subcategory “explosion” can be caused by equipment malfunctions, human errors or other sources, and a new potential main category is suggested for such cases. The explanation provided by ChatGPT suggests that the result is caused by the vague subcategory “explosion,” which does not provide enough information for ChatGPT to determine the meaning of the explosion. In this case of insufficient information, human can review the original data to confirm the “explosion” specifically refers to “equipment explosion,” and label “explosion” by the existing main category “equipment & infrastructure issues.” However, due to limited input message, all original data cannot be provided along with subcategories. In the absence of original data, ChatGPT’s response is evaluated as comprehensive and reasonable. The evaluation result suggests that ChatGPT is capable of performing the data saturation test and achieving results comparable to humans.

Performance Summary

In summary, the experimental results show that the proposed approach and crafted prompt patterns can effectively guide ChatGPT to perform GT. ChatGPT’s performance in each step of the proposed approach is summarized as follows.

In the information extraction step, ChatGPT demonstrates human-comparable capability in extracting causal information. This capability is more stable in the one-shot scenario compared to the zero-shot scenario but decreases following the increasing number of cases processed by ChatGPT each time.

In the open coding step, ChatGPT demonstrates human-comparable capability in generating subcategories in all scenarios. ChatGPT-generated subcategories can capture the meanings of all human-generated subcategories with high similarity.

In the axial coding step, ChatGPT demonstrates an excellent capability in generating main categories which capture the meanings of human-generated main categories. However, ChatGPT tends to generate broader main categories compared with humans.

In the selective coding step, ChatGPT is capable of correctly grasping the topic of the research data and generating appropriate and diverse core categories in different scenarios, which provides alternative perspectives for exploring research data. Furthermore, ChatGPT can provide comprehensive explanations for the relationships between categories, which

facilitates in generating the risk terminology and taxonomy, and constructing the risk picture.

In the data saturation test, ChatGPT is capable of performing the data saturation test and achieving human-like results in most scenarios. For the results inconsistent with human, ChatGPT can provide thoughtful and reasonable explanations.

All the aforementioned results suggest that ChatGPT does have the potential to perform GT. By adding an example to a prompt or providing less cases to be processed in a prompt, the results generated by ChatGPT could be improved. Additionally, incorporating ChatGPT as a research assistant to humans applying GT can reduce the time and efforts, and enable more efficient information extraction and data coding. Therefore, ChatGPT is a promising new alternative to assist performing GT.

Discussion

This research offers contributions to both theoretical and practical aspects.

Theoretical Contributions

From the theoretical aspect, three innovations in research methodology are presented in this research.

We propose a novel idea of utilizing ChatGPT to perform GT to do qualitative data analysis and develop a ChatGPT-performed GT approach. To the best of our knowledge, this is the first research that integrates ChatGPT and GT, which not only makes the data coding process more efficient and reliable, but also can be utilized by researchers without text mining knowledge. Our approach is flexibly adaptable to various qualitative data analysis tasks.

We design well-structured and generalized prompt patterns to guide ChatGPT to perform GT. The experiment results have verified that our crafted prompt patterns effectively elicit the desired results from ChatGPT. These prompt patterns create a standardized approach for humans to interact with ChatGPT, making it possible to conduct GT without text mining knowledge.

We propose general quantitative metrics to evaluate ChatGPT's performance in each step of our approach. The experiments have verified that our metrics can effectively measure ChatGPT's performance, which provides comprehensive perspectives to compare and understand the similarities and differences between ChatGPT and humans.

Practical Contributions

From the practical aspect, this research demonstrates how to guide ChatGPT to perform required tasks, rigorously evaluate outputs, and ensure high-quality results.

We demonstrated that to guide ChatGPT to perform required tasks, we need to use prompt to tell ChatGPT what role it will play, what data it will use, what analysis it will perform, and what format it will follow to generate required output. It is the same way to assign a research assistant to do the job for you.

We illustrate that it is important to set criteria to evaluate the quality of the output generated by ChatGPT. The same as a human research assistant, ChatGPT will also make mistakes. We need to check if the answer provided by ChatGPT makes sense. It is our responsibility to control the quality of the work of our research assistant.

We introduce ChatGPT as a research assistant to assist researchers to do qualitative analysis. ChatGPT performs tedious and time-consuming work such as information extraction and data coding, while humans are responsible for supervising ChatGPT's work and verifying results. This greatly promotes the efficiency and quality of GT research, and as a result contributes to theory building and development in IS research. Researchers can develop a variety of ways using ChatGPT or other generative AI tools for the theoretical conceptualization works, and our approach is just one of the efforts.

Limitations and Future Research

Some limitations can be addressed in future research. First, this research uses the results provided by domain experts as a comparison benchmark to assess the performance of ChatGPT. However, the experts' opinions cannot guarantee the true quality of the assessment of the analysis. Due to the absence of existing validation methods in GT, further research is needed to explore more advanced validation methods [52, 80] and the possible collaboration between human and AI [96, 99] to improve the performance of our approach.

Second, the term “human-comparable” in this research is used to describe ChatGPT achieving specific capabilities measured by the metrics in specific tasks comparable to humans under experimental conditions. While metrics show success in specific tasks, such as information extraction and open coding, other aspects, like axial coding, remain challenging. Further validation with alternative Generative AI Models or customized ChatGPT may provide a more consistent evaluation of “human-comparable” performance.

Third, the validation of our proposed approach is limited to the current version of ChatGPT-4 and text data. Recently, OpenAI has released a new version called GPT-4o or GPT-4o, which offers substantial improvements over its predecessors by introducing multimodal capabilities, larger context windows, efficient tokenization, and faster processing speeds, achieving state-of-the-art performance in text, audio, video, and image generation and understanding [59]. Google developed Gemini, a family of highly capable multimodal models [41]. Gemini and ChatGPT showcase distinct strengths across various performance metrics. Gemini's integration with Google Search provides a notable advantage in actual accuracy, whereas ChatGPT excels in conversational fluidity and creative expression [93]. Other generative AI systems, such as Microsoft Copilot [60] and Anthropic Claude AI [1], also exist. With the new development of LLMs, future research can focus on experimenting with alternative LLMs in diverse scenarios to enhance the performance and scalability of our approach. A longitudinal research framework to assess performance over time with evolving AI models and across various data types would strengthen the replicability and generalizability of our results.

Fourth, the generalizability of our approach across a variety of domains needs further validation. For example, GT is suggested to conduct rigorous literature reviews [117, 124], and we need to verify if the ChatGPT-performed GT approach can also be applied and be effective in such work. In addition, this research mainly focuses on the use of ChatGPT for

qualitative analysis. We may further explore how Generative AI can be used for combining both quantitative and qualitative research. For example, in financial risk analysis, Generative AI can be applied to examining the consistency between quantitative financial ratios and qualitative narrative disclosures in financial reports, or to integrate quantitative and qualitative criteria for credit risk assessment [18, 50].

Conclusion

This research investigates ChatGPT's potential to perform grounded theory approach to do risk analysis and provides a detailed understanding of ChatGPT's capability on this task. The ChatGPT-performed GT approach is proposed to guide ChatGPT in the form of prompt patterns. The quantitative metrics are designed to evaluate ChatGPT's performance in comparison with manual GT. The experimental results show that the proposed approach and crafted prompt patterns can effectively guide ChatGPT to perform GT well in each step of the process. Incorporating ChatGPT as a research assistant to perform GT can reduce the time and mental effort required, enabling more efficient information extraction and data coding. With the rapid advances in generative AI, such as GPT-4o, Gemini, and Co-pilot, it is important to study how to facilitate effective collaboration between humans and AI, and how to evaluate and control the quality of research.

Disclosure statement

The authors have no conflicts of interest to disclose.

Funding

This work was supported by the Natural Sciences and Engineering Research Council of Canada [2018-06743].

Notes on contributors

Yaxian Zhou (yaxianaimhigh@gmail.com) is a PhD candidate in Management Science at Dalian University of Technology, China, and a visiting student of Information Systems at McMaster University, Canada. Her research interests are in the areas of grounded theory, risk management, and energy management. Her research has been published in such journals as *Energy and Buildings*, *Frontiers of Engineering Management*, and others.

Yufei Yuan (yuanyuf@mcmaster.ca; corresponding author) is a Professor of Information Systems at DeGroote School of Business, McMaster University, Canada. He received his PhD in computer information systems from the University of Michigan. Dr. Yuan's research interests include artificial intelligence, big data analytics, security and privacy, mobile commerce, emergency response systems, web-based negotiation support systems, and information systems in health care. He has published more than 100 papers in such journals as *Management Science*, *Journal of Management Information Systems*, *MIS Quarterly*, *Decision Support Systems*, *Information & Management*, *European Journal of Information Systems*, *European Journal of Operational Research*, *Communications of the ACM*, *IEEE Security and Privacy*, and many others.

Kai Huang (khuang@mcmaster.ca) is a Professor of Operations Management at McMaster University, Canada. He received his PhD from Georgia Institute of Technology. Dr. Huang's research

interests include the optimization under uncertainty and the data-driven optimization techniques with applications in business analytics and supply chain management. His research has been published in journals such as *Operations Research*, *Mathematical Programming*, *Naval Research Logistics*, *European Journal of Operational Research*, *Transportation Research Part E: Logistics and Transportation Review*, and many others.

Xiangpei Hu (drhxp@dlut.edu.cn) is a Distinguished Professor of Management Science in School of Economics and Management at Dalian University of Technology. He received his PhD in Management Science from Harbin Institute of Technology, China. Dr. Hu's research interests are in the areas of operations research, electronic commerce, and logistics management. His research has been published in such journals as *European Journal of Operational Research*, *Computers & Operations Research*, *Decision Support Systems*, *Annals of Operations Research*, *Transportation Research Part E: Logistics and Transportation Review*, *IEEE Internet of Things Journal*, and many others.

ORCID

Yufei Yuan  <http://orcid.org/0000-0002-8388-8433>

References

1. Adetayo, A.J.; Aborisade, M.O.; and Sanni, B.A. Microsoft Copilot and Anthropic Claude AI in education and library service. *Library Hi Tech News*, ahead-of-print, (2024).
2. Alawida, M.; Mejri, S.; Mehmood, A.; Chikhaoui, B.; and Abiodun, O.I. A comprehensive study of ChatGPT: Advancements, limitations, and ethical considerations in natural language processing and cybersecurity. *Information*, 14, 8 (2023), 462.
3. Alhelou, H.H.; Hamedani-Golshan, M.E.; Njenda, T.C.; and Siano, P. A survey on power system blackout and cascading events: Research motivations and challenges. *Energies*, 12, 4 (2019), 682.
4. Apramian, T.; Cristancho, S.; Watling, C.; and Lingard, L. (Re)Grounding grounded theory: A close reading of theory in four schools. *Qualitative Research*, 17, 4 (2017), 359–376.
5. Aslipour, H.; and Zargar, M.R. Developing grounded theory systematic approach for public policy researches. *International Journal of Qualitative Methods*, 21, (2022), 16094069221090357.
6. Aven, T. Risk Management. In T. Aven (ed.), *Risk Analysis*. New York: John Wiley & Sons, 2015, pp. 4–10.
7. Aven, T. An emerging new risk analysis science: Foundations and implications. *Risk Analysis*, 38, 5 (2018), 876–888.
8. Baggott, S.S., and Santos, J.R. A risk analysis framework for cyber security and critical infrastructure protection of the U.S. electric power grid. *Risk Analysis*, 40, 9 (2020), 1744–1761.
9. Bahrini, A.; Khamoshifar, M.; Abbasimehr, H.; Riggs, R.J.; Esmaceli, M.; Majdabadkohne, R.M.; and Pasehvar, M. ChatGPT: Applications, opportunities, and threats. In *2023 Systems and Information Engineering Design Symposium (SIEDS)*. Charlottesville: IEEE, 2023, pp. 274–279.
10. Bang, Y.; Cahyawijaya, S.; Lee, N.; Dai, W.; Su, D.; Wilie, B.; Lovenia, H.; Ji, Z.; Yu, T.; Chung, W.; A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*. Bali: Association for Computational Linguistics, 2023, pp. 675–718.
11. Berente, N.; Seidel, S.; and Safadi, H. Research commentary-data-driven computationally intensive theory development. *Information Systems Research*, 30, 1 (2019), 50–64.
12. Berthelsen, C.B.; Grimshaw-Aagaard, S.; and Hansen, C. Developing a guideline for reporting and evaluating grounded theory research studies. *International Journal of Health Sciences*, 6, 1 (2018), 13.

13. Birks, D.F.; Fernandez, W.; Levina, N.; and Nasirin, S. Grounded theory method in information systems research: Its nature, diversity and opportunities. *European Journal of Information Systems*, 22, 1 (2013), 1–8.
14. Brillinger, M.; Henze, J.; Albert, C.; and Schwarze, R. Integrating nature-based solutions in flood risk management plans: A matter of individual beliefs? *Science of The Total Environment*, 795, (2021), 148896.
15. Bubeck, S.; Chandrasekaran, V.; Eldan, R.; Gehrke, J.; Horvitz, E.; Kamar, E.; Lee, P.; Lee, Y.T.; Li, Y.; Lundberg, S.; and Nori, H. Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712*, (2023).
16. Budhwar, P.; Chowdhury, S.; Wood, G.; Aguinis, H.; Bamber, G.J.; Beltran, J.R.; Boselie, P.; Lee Cooke, F.; Decker, S.; DeNisi, A.; and Dey, P.K., Human resource management in the age of generative artificial intelligence: Perspectives and research directions on ChatGPT. *Human Resource Management Journal*, 33, 3 (2023), 606–659.
17. Burton-Jones, A.; Butler, B.S.; Scott, S.; and Xu, S.X. Next-generation information systems theorizing: A call to action. *MIS Quarterly*, 45, 1 (2021), 301–314.
18. Cao, L. AI in finance: Challenges, techniques, and opportunities. *ACM Computing Surveys*, 55, 3 (2022), 64:1–64:38.
19. Cascella, M.; Montomoli, J.; Bellini, V.; and Bignami, E. Evaluating the feasibility of ChatGPT in healthcare: An analysis of multiple clinical and research scenarios. *Journal of Medical Systems*, 47, 1 (2023), 33.
20. Chametzky, B. Coding in classic grounded theory: I’ve done an interview; Now what? *Sociology Mind*, 6, 4 (2016), 163–172.
21. Charmaz, K. Grounded theory: Objectivist and constructivist methods. *Handbook of Qualitative Research*, 2, 1 (2000), 509–535.
22. Charmaz, K. *Constructing Grounded Theory: A Practical Guide through Qualitative Analysis*. Thousand Oaks: Sage Publications Ltd, 2006.
23. Charmaz, K. “With constructivist grounded theory you can’t hide”: Social justice research and critical inquiry in the public sphere. *Qualitative Inquiry*, 26, 2 (2020), 165–176.
24. Charmaz, K. and Thornberg, R. The pursuit of quality in grounded theory. *Qualitative Research in Psychology*, 18, 3 (2021), 305–327.
25. Chen, J.; Chen, P.; and Wu, X. Generating Chinese event extraction method based on ChatGPT and prompt learning. *Applied Sciences*, 13, 17 (2023), 9500.
26. Chen, W.; Lu, Y.; Qiu, L.; and Kumar, S. Designing personalized treatment plans for breast cancer. *Information Systems Research*, 32, 3 (2021), 932–949.
27. Chen, X.; Zhang, X.; Liu, Y.; Wang, Z.; Zhou, Y.; and Chu, M. RISK-GPT: Using ChatGPT to construct a reliable risk factor database for all known diseases. *Journal of Global Health*, 13, (2023), 03037.
28. Clarke, A.E. *Situational Analysis: Grounded Theory After the Postmodern Turn*. Thousand Oaks: Sage Publications Ltd, 2005.
29. Clarke, A.E. Grounded theory: Critiques, debates, and situational analysis. *The SAGE handbook of social science methodology*, (2007), 423–442.
30. Corbin, J.M.; and Strauss, A.L. Grounded theory research: Procedures, canons, and evaluative criteria. *Qualitative Sociology*, 13, 1 (1990), 3–21.
31. Cuesta-Benjumea, C.D.L.; Lidon-Cerezuela, B.; Abad-Corpa, E.; Meseguer-Liza, C.; and Arredondo-Gonzalez, C.P. Managing and keeping control: A qualitative synthesis of nursing and care staff strategies to prevent older people from falling. *Journal of Advanced Nursing*, 77, 7 (2021), 3008–3019.
32. Darkhabani, M.; Alrifaii, M.A.; Elsalti, A.; Dvir, Y.M.; and Mahroum, N. ChatGPT and autoimmunity – A new weapon in the battlefield of knowledge. *Autoimmunity Reviews*, 22, 8 (2023), 103360.
33. Deng, Y.; Zhang, Y.; Yuan, Z.; Li, R.Y.M.; and Gu, T. Analyzing subway operation accidents causations: Apriori algorithm and network approaches. *International Journal of Environmental Research and Public Health*, 20, 4 (2023), 3386.

34. Derakhshanfar, H.; Ochoa, J.J.; Kirytopoulos, K.; Mayer, W.; and Tam, V.W.Y. Construction delay risk taxonomy, associations and regional contexts: A systematic review and meta-analysis. *Engineering, Construction and Architectural Management*, 26, 10 (2019), 2364–2388.
35. van Dis, E.A.M.; Bollen, J.; Zuidema, W.; van Rooij, R.; and Bockting, C.L. ChatGPT: Five priorities for research. *Nature*, 614, 7947 (2023), 224–226.
36. Dowling, M.; and Lucey, B. ChatGPT for (Finance) research: The bananarama conjecture. *Finance Research Letters*, 53, (2023), 103662.
37. Dwivedi, Y.K.; Kshetri, N.; Hughes, L.; Slade, E.L.; Jeyaraj, A.; Kar, A.K.; Baabdullah, A.M.; Koohang, A.; Raghavan, V.; Ahuja, M.; and Albanna, H. “So what if ChatGPT wrote it?” Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, 71, (2023), 102642.
38. Feuerriegel, S.; Hartmann, J.; Janiesch, C.; and Zschech, P. Generative AI. *Business & Information Systems Engineering*, 66, 1 (2024), 111–126.
39. Frederico, G.F. ChatGPT in supply chains: Initial evidence of applications and potential research agenda. *Logistics*, 7, 2 (2023), 26.
40. Fu, L.; Yang, Q.; Liu, X.; and He, L. Risk assessment of infectious disease epidemic based on fuzzy Bayesian network. *Risk Analysis*, 44, 1 (2023), 40–53.
41. Team, G.; Anil, R.; Borgeaud, S.; Alayrac, J.B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A.M.; Hauth, A.; Millican, K.; and Silver, D. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, (2024).
42. George, J.; Gerhart, N.; and Torres, R. Uncovering the truth about fake news: A research model grounded in multi-disciplinary literature. *Journal of Management Information Systems*, 38, 4 (2021), 1067–1094.
43. Gill, S.S.; and Kaur, R. ChatGPT: Vision and challenges. *Internet of Things and Cyber-Physical Systems*, 3, (2023), 262–271.
44. Glaser, B.G. *Basics of Grounded Theory Analysis: Emergence vs. Forcing*. Mill Valley: Sociology Press, 1992.
45. Glaser, B.G. *The Grounded Theory Perspective: Conceptualization Contrasted with Description*. Mill Valley: Sociology Press, 2001.
46. Glaser, B.G. Conceptualization: On theory and theorizing using grounded theory. *International Journal of Qualitative Methods*, 1, 2 (2002), 23–38.
47. Glaser, B.G. and Holton, J. Remodeling grounded theory. *Forum: Qualitative Social Research*, 5, 2 (2004), Art. 4.
48. Glaser, B.G.; and Strauss, A.L. *The Discovery of Grounded Theory: Strategies for Qualitative Research*. New Brunswick: Aldine Transaction, 1967.
49. Glickman, M., and Zhang, Y. AI and generative AI for research discovery and summarization. *Harvard Data Science Review*, 6, 2 (2024).
50. Goodell, J.W.; Kumar, S.; Lim, W.M.; and Pattnaik, D. Artificial intelligence and machine learning in finance: Identifying foundations, themes, and research clusters from bibliometric analysis. *Journal of Behavioral and Experimental Finance*, 32, (2021), 100577.
51. Google AI for Developers. Gemini models. 2024. <https://ai.google.dev/models/gemini> (accessed on April 11, 2024).
52. Grossmann, I.; Feinberg, M.; Parker, D.C.; Christakis, N.A.; Tetlock, P.E.; and Cunningham, W.A. AI and the transformation of social science research. *Science*, 380, 6650 (2023), 1108–1109.
53. Guikema, S. Artificial intelligence for natural hazards risk analysis: Potential, challenges, and research needs. *Risk Analysis*, 40, 6 (2020), 1117–1123.
54. Hamilton, L.; Elliott, D.; Quick, A.; Smith, S.; and Choplin, V. Exploring the use of AI in qualitative analysis: A comparative study of guaranteed income data. *International Journal of Qualitative Methods*, 22, (2023), 16094069231201504.

55. Han, R.; Peng, T.; Yang, C.; Wang, B.; Liu, L.; and Wan, X. Is information extraction solved by ChatGPT? An analysis of performance, evaluation criteria, robustness and errors. *arXiv preprint arXiv:2305.14450*, (2023).
56. He, L.; Ma, X.; Qiao, W.; and Yang, L. A methodology to assess the causation relationship of seafarers' unsafe acts for ship grounding accidents based on Bayesian SEM. *Ocean & Coastal Management*, 225, (2022), 106189.
57. Heath, H.; and Cowley, S. Developing a grounded theory approach: A comparison of Glaser and Strauss. *International Journal of Nursing Studies*, 41, 2 (2004), 141–150.
58. Huang, T.; Baiocchi, M.; and Lei, X. An ontological approach for automatic tracking causes of blackouts in power systems. In *2022 IEEE 21st Mediterranean Electrotechnical Conference (MELECON)*. Palermo: IEEE, 2022, pp. 813–818.
59. Islam, R., and Moushi, O.M. GPT-4o: The cutting-edge advancement in multimodal LLM. *Authorea Preprints*, (2024).
60. Khan, A. Microsoft Copilot Studio. In A. Khan (ed.), *Introducing Microsoft Copilot for Managers: Enhance Your Team's Productivity and Creativity with Generative AI-Powered Assistant*. Berkeley: Apress, 2024, pp. 621–694.
61. Khan, A.; Qu, X.; and Madzikanda, B. An exploratory study on risk identification of cross-boundary innovation of manufacturing enterprises based on grounded theory. *Creativity and Innovation Management*, 31, 3 (2022), 492–508.
62. Kıcıman, E.; Ness, R.; Sharma, A.; and Tan, C. Causal reasoning and large language models: Opening a new frontier for causality. *arXiv preprint arXiv:2305.00050*, (2023).
63. Konecki, K.T. Visual grounded theory: A methodological outline and examples from empirical work. *Journal of Sociology*, (2011), 131–160.
64. Konecki, K.T. Classic grounded theory—The latest version: Interpretation of classic grounded theory as a meta-theory for research. *Symbolic Interaction*, 41, 4 (2018), 547–564.
65. Krzysztof, S.; and Złotecka, D. The risk of large blackout failures in power systems. *Archives of Electrical Engineering*, (2019), 411–426.
66. Lande, D., and Strashnoy, L. Causality network formation with ChatGPT. *Available at SSRN 4464477*, (2023).
67. Le, P.T.; Kirytopoulos, K.; Chileshe, N.; and Rameezdeen, R. Taxonomy of risks in PPP transportation projects: A systematic literature review. *International Journal of Construction Management*, 22, 2 (2022), 166–181.
68. Levina, N. All information systems theory is grounded theory. *MIS Quarterly*, 45, 1 (2021), 489–494.
69. Levkovich, I.; and Elyoseph, Z. Suicide risk assessments through the eyes of ChatGPT-3.5 versus ChatGPT-4: Vignette study. *JMIR Mental Health*, 10, 1 (2023), e51232.
70. Li, B.; Fang, G.; Yang, Y.; Wang, Q.; Ye, W.; Zhao, W. and Zhang, S., Evaluating ChatGPT's information extraction capabilities: An assessment of performance, explainability, calibration, and faithfulness. *arXiv preprint arXiv:2304.11633*, (2023).
71. Li, X.; Jiang, H.; and Liang, X. Risk transmission mechanism of domestic cluster epidemic caused by overseas imported cases: Multiple case studies based on grounded theory. *International Journal of Environmental Research and Public Health*, 19, 18 (2022), 11810.
72. Li, X.; Jiang, H.; and Liang, X. Early stage risk identification and governance of major emerging infectious diseases: A double-case study based on the Chinese context. *Risk Management and Healthcare Policy*, 16, (2023), 635–653.
73. Liu, W.; Zhang, J.; Shi, Y.; Lee, P.T.-W.; and Liang, Y. Intelligent logistics transformation problems in efficient commodity distribution. *Transportation Research Part E: Logistics and Transportation Review*, 163, (2022), 102735.
74. Lumivero. An overview of grounded theory qualitative research. 2020. <https://lumivero.com/resources/an-overview-of-grounded-theory-qualitative-research/> (accessed on September 24, 2023).
75. Ma, J.; Shuang, F.; Guo, W.; Du, Q.; Li, J.; and Zheng, H. Application of multi-factor dynamic interaction graph in vulnerability assessment and online monitoring of transmission lines. *International Journal of Electrical Power & Energy Systems*, 143, (2022), 108435.

76. Mao, R.; Chen, G.; Zhang, X.; Guerin, F.; and Cambria, E. GPTEval: A survey on assessments of ChatGPT and GPT-4. *arXiv preprint arXiv:2308.12488*, (2023).
77. Maruping, L.M.; Venkatesh, V.; Thong, J.Y.L.; and Zhang, X. A risk mitigation framework for information technology projects: A cultural contingency perspective. *Journal of Management Information Systems*, 36, 1 (2019), 120–157.
78. Matavire, R.; and Brown, I. Profiling grounded theory approaches in information systems research. *European Journal of Information Systems*, 22, 1 (2013), 119–129.
79. Mills, J.; Bonner, A.; and Francis, K. The development of constructivist grounded theory. *International Journal of Qualitative Methods*, 5, 1 (2006), 25–35.
80. Moor, M.; Banerjee, O.; Abad, Z.S.H.; Krumholz, H.M.; Leskovec, J.; Topol, E.J. and Rajpurkar, P. Foundation models for generalist medical artificial intelligence. *Nature*, 616, 7956 (2023), 259–265.
81. Morgan, D.L. Exploring the use of artificial intelligence for qualitative data analysis: The case of ChatGPT. *International Journal of Qualitative Methods*, 22, (2023), 16094069231211248.
82. Müller, B.; and Olbrich, S. Developing Theories in Information Systems Research: The Grounded Theory Method Applied. In Y.K. Dwivedi, M.R. Wade and S.L. Schneberger (eds). *Information Systems Theory: Explaining and Predicting Our Digital Society*. New York: Springer, 2012, pp. 323–347.
83. Wiesche, M.; Jurisch, M.C.; Yetton, P.W. and Krcmar, H. Grounded theory methodology in information systems research. *MIS Quarterly*, 41, 3 (2017), 685–701.
84. Nelson, L.K. Computational grounded theory: A methodological framework. *Sociological Methods & Research*, 49, 1 (2020), 3–42.
85. Nelson, L.K.; Burk, D.; Knudsen, M.; and McCall, L. The future of coding: A comparison of hand-coding and three types of computer-assisted text analysis methods. *Sociological Methods & Research*, 50, 1 (2021), 202–237.
86. Nyqvist, R.; Peltokorpi, A.; and Seppänen, O. Can ChatGPT exceed humans in construction project risk management? *Engineering, Construction and Architectural Management*, 31, 13 (2024), 223–243.
87. Odacioglu, E.C., and Zhang, L. Text mining for rendering theory: Integrating topic modeling to grounded theory. Available at SSRN 4141372, (2022).
88. OpenAI. OpenAI platform. 2023. <https://platform.openai.com> (accessed on September 28, 2023).
89. OpenAI. GPT-4. 2023. <https://openai.com/research/gpt-4> (accessed on September 28, 2023).
90. OpenAI. ChatGPT. 2023. <https://chat.openai.com> (accessed on September 30, 2023).
91. Piriyaikul, I. and Piriyaikul, R. Automated analysis of causal relationships in customer reviews. *Research Square Preprint*, (2023).
92. R OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, (2024).
93. Rane, N.; Saurabh, C.; and Rane, J. Gemini versus ChatGPT: Applications, performance, architecture, capabilities, and implementation. *Journal of Applied Artificial Intelligence*, 5, 1 (2024), 69–93.
94. Ray, P.P. ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, 3, (2023), 121–154.
95. Redman-MacLaren, M.; and Mills, J. Transformational grounded theory: Theory, voice, and action. *International Journal of Qualitative Methods*, 14, 3 (2015), 1–12.
96. Revilla, E.; Saenz, M.J.; Seifert, M.; and Ma, Y. Human–Artificial Intelligence collaboration in prediction: A field experiment in the retail industry. *Journal of Management Information Systems*, 40, 4 (2023), 1071–1098.
97. Roumeliotis, K.I.; and Tselikas, N.D. ChatGPT and Open-AI models: A preliminary review. *Future Internet*, 15, 6 (2023), 192.
98. Sahoo, P.; Singh, A.K.; Saha, S.; Jain, V.; Mondal, S.; and Chadha, A. A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*, (2024).

99. Sharma, A.; Lin, I.W.; Miner, A.S.; Atkins, D.C.; and Althoff, T. Human–AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nature Machine Intelligence*, 5, 1 (2023), 46–57.
100. Shi, D.; Guan, J.; Zurada, J.; and Manikas, A. A data-mining approach to identification of risk factors in safety management systems. *Journal of Management Information Systems*, 34, 4 (2017), 1054–1081.
101. Shield, S.A.; Quiring, S.M.; Pino, J.V.; and Buckstaff, K. Major impacts of weather events on the electrical power delivery system in the United States. *Energy*, 218, (2021), 119434.
102. Singh, H.; and Singh, A. ChatGPT: Systematic review, applications, and agenda for multi-disciplinary research. *Journal of Chinese Economic and Business Studies*, 21, 2 (2023), 193–212.
103. Stokel-Walker, C. and Noorden, R.V. What ChatGPT and generative AI mean for science. *Nature*, 614, 7947 (2023), 214–216.
104. Strauss, A.L.; and Corbin, J.M. *Basics of Qualitative Research: Grounded Theory Procedures and Techniques*. Thousand Oaks: Sage Publications Inc, 1990.
105. Strauss, A.L.; and Corbin, J.M. Grounded theory methodology: An overview. In N.K. Denzin and Y.S. Lincoln (eds.), *Handbook of Qualitative Research*. Thousand Oaks: Sage Publications Inc, 1994, pp. 273–285.
106. Strauss, A.L.; and Corbin, J.M. *Grounded Theory in Practice*. Thousand Oaks: Sage Publications Inc, 1997.
107. Susarla, A.; Gopal, R.; Thatcher, J.B.; and Sarker, S. The Janus effect of Generative AI: Charting the path for responsible conduct of scholarly activities in information systems. *Information Systems Research*, 34, 2 (2023), 399–408.
108. Tamburri, D.A.; Palomba, F.; and Kazman, R. Success and failure in software engineering: A followup systematic literature review. *IEEE Transactions on Engineering Management*, 68, 2 (2021), 599–611.
109. Tang, X.; Wang, M.; Wang, Q.; Zhang, J.; Li, H.; and Tang, J. Exploring technical decision-making risks in construction megaprojects using grounded theory and system dynamics. *Computational Intelligence and Neuroscience*, 2022, 1 (2022), 9598781.
110. The University of Texas at Austin. *The Timeline and Events of the February 2021 Texas Electric Grid Blackouts*. The University of Texas at Austin. 2021. <https://energy.utexas.edu/sites/default/files/UTAustin%20%282021%29%20EventsFebruary2021TexasBlackout%2020210714.pdf>.
111. Themelis, C.; Sime, J.-A.; and Thornberg, R. Informed grounded theory: A symbiosis of philosophy, methodology, and art. *Scandinavian Journal of Educational Research*, 67, 7 (2023), 1086–1099.
112. Tian, S.; Jin, Q.; Yeganova, L.; Lai, P.T.; Zhu, Q.; Chen, X.; Yang, Y.; Chen, Q.; Kim, W.; Comeau, D.C.; and Islamaj, R. Opportunities and challenges for ChatGPT and large language models in biomedicine and health. *Briefings in Bioinformatics*, 25, 1 (2024), bbad493.
113. Tu, R.; Ma, C.; and Zhang, C. Causal-discovery performance of ChatGPT in the context of neuropathic pain diagnosis. *arXiv preprint arXiv:2301.13819*, (2023).
114. Urquhart, C.; and Fernández, W. Using grounded theory method in information systems: The researcher as blank slate and other myths. *Journal of Information Technology*, 28, 3 (2013), 224–236.
115. Urquhart, C.; Lehmann, H.; and Myers, M.D. Putting the “theory” back into grounded theory: Guidelines for grounded theory studies in information systems. *Information Systems Journal*, 20, 4 (2010), 357–381.
116. U.S. Department of State. Risk analysis. 2023. <https://www.state.gov/risk-analysis/> (accessed on October 2, 2023).
117. Walsh, I.; and Rowe, F. BIBGT: combining bibliometrics and grounded theory to conduct a literature review. *European Journal of Information Systems*, 32, 4 (2023), 653–674.
118. Wang, T.; Kannan, K.N.; and Ulmer, J.R. The association between the disclosure and the realization of information security risk factors. *Information Systems Research*, 24, 2 (2013), 201–218.

119. Wang, Z.; Jiang, C.; Zhao, H.; and Ding, Y. Mining semantic soft factors for credit risk evaluation in Peer-to-Peer lending. *Journal of Management Information Systems*, 37, 1 (2020), 282–308.
120. Watters, C.; and Lemanski, M.K. Universal skepticism of ChatGPT: A review of early literature on chat generative pre-trained transformer. *Frontiers in Big Data*, 6, (2023), 1224976.
121. Way, S.; and Yuan, Y. A framework for collaborative disaster response: A grounded theory approach. In M. Schoop and D.M. Kilgour (eds.), *Group Decision and Negotiation. A Socio-Technical Perspective*. Cham: Springer International Publishing, 2017, pp. 33–46.
122. Wei, X.; Cui, X.; Cheng, N.; Wang, X.; Zhang, X.; Huang, S.; Xie, P.; Xu, J.; Chen, Y.; Zhang, M.; Jiang, Y.; and Han, W. ChatIE: Zero-shot information extraction via chatting with ChatGPT. *arXiv preprint arXiv:2302.10205*, (2023).
123. White, J.; Fu, Q.; Hays, S.; Sandborn, M.; Olea, C.; Gilbert, H.; Elnashar, A.; Spencer-Smith, J.; and Schmidt, D.C. A prompt pattern catalog to enhance prompt engineering with ChatGPT. *arXiv preprint arXiv:2302.11382*, (2023).
124. Wolfswinkel, J.F.; Furtmueller, E.; and Wilderom, C.P.M. Using grounded theory as a method for rigorously reviewing literature. *European Journal of Information Systems*, 22, 1 (2013), 45–55.
125. Wu, T.; He, S.; Liu, J.; Sun, S.; Liu, K.; Han, Q.L.; and Tang, Y. A brief overview of ChatGPT: The history, status quo and potential future development. *IEEE-CAA Journal of Automatica Sinica*, 10, 5 (2023), 1122–1136.
126. Xiong, Y.; Qi, H.; Li, Z.; and Zhang, Q. Where risk, where capability? Building the emergency management capability structure of coal mining enterprises based on risk matching perspective. *Resources Policy*, 83, (2023), 103695.
127. Xu, R.; Luo, F.; Chen, G.; Zhou, F.; and Abdulahi, E.W. Application of HFACS and grounded theory for identifying risk factors of air traffic controllers' unsafe acts. *International Journal of Industrial Ergonomics*, 86, (2021), 103228.
128. Xue, Z.; Xu, C.; and Xu, X. Application of ChatGPT in natural disaster prevention and reduction. *Natural Hazards Research*, 3, 3 (2023), 556–562.
129. Yadav, D.; Zhang, S.; Jin, T.; Krishnan, P.; and Clarke, D. Generative AI based virtual assistant for reconciliation research. *Amazon Science*, 2024. <https://www.amazon.science/publications/generative-ai-based-virtual-assistant-for-reconciliation-research> (accessed on September 1, 2024).
130. Yang, J.; Han, S.C.; and Poon, J. A survey on extraction of causal relations from natural language text. *Knowledge and Information Systems*, 64, 5 (2022), 1161–1186.
131. Yang, L.; Wang, X.; Zhu, J.; and Qin, Z. Risk factors identification of unsafe acts in deep coal mine workers based on grounded theory and HFACS. *Frontiers in Public Health*, 10, (2022), 852612.
132. Yang, S.; Chen, W.; and Zhang, X. Heterogeneous evolution of power system vulnerability in cascading failure graphs. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 69, 1 (2022), 179–183.
133. Yang, Y.; Wang, Y.; Easa, S.M.; and Yan, X. Factors affecting road tunnel construction accidents in China based on grounded theory and DEMATEL. *International Journal of Environmental Research and Public Health*, 19, 24 (2022), 16677.
134. Yu, Q.; Cao, N.; Liu, Q.; Qu, Y.; and Zhang, Y. Self-organized criticality and trend analysis in time series of blackouts for the China power grid. *Mathematical Problems in Engineering*, 2020, 1 (2020), 3075935.
135. Yuan, C.; Xie, Q.; and Ananiadou, S. Zero-shot temporal relation extraction with ChatGPT. *arXiv preprint arXiv:2304.05454*, (2023).
136. Zhao, X.; Peng, B.; Zheng, C.; and Wan, A. Business model innovation risk factors based on grounded theory: A multiple-case analysis of cold chain logistics companies in China. *Managerial and Decision Economics*, 43, 6 (2022), 2108–2118.

137. Zhao, Y.; Chen, W.; Yang, Z.; Li, Z.; and Wang, Y. Analysis on risk factors related delay in PCPs. *Engineering, Construction and Architectural Management*, 30, 10 (2023), 4609–4644.
138. Zhu, G.; Chen, G.; Zhu, J.; Meng, X.; and Li, X. Modeling the evolution of major storm-disaster-induced accidents in the offshore oil and gas industry. *International Journal of Environmental Research and Public Health*, 19, 12 (2022), 7216.
139. Zhu, Y.; Wang, X.; Chen, J.; Qiao, S.; Ou, Y.; Yao, Y.; Deng, S.; Chen, H.; and Zhang, N. LLMs for knowledge graph construction and reasoning: Recent capabilities and future opportunities. *arXiv preprint arXiv:2305.13168*, (2023).
140. Zhu, Y.; Yuan, H.; Wang, S.; Liu, J.; Liu, W.; Deng, C.; Chen, H.; Dou, Z.; and Wen, J.R. Large language models for information retrieval: A survey. *arXiv preprint arXiv:2308.07107*, (2023).
141. Zhu, Y.; Zhang, P.; Haq, E.-U.; Hui, P.; and Tyson, G. Can ChatGPT reproduce human-generated labels? A study of social computing tasks. *arXiv preprint arXiv:2304.10145*, (2023).
142. Zupanic, M.A.; Schafft, H.-A.; Pieper, R.; Lindemann, A.-K.; and Mader, A. A conceptual framework for the identification of food safety risks in global commodity flows exemplified by agricultural bulk commodities. *Operations and Supply Chain Management-An International Journal*, 15, 1 (2022), 79–92.