Contents

Efficient Development of a Domain-Specific Question Answering System for Mining Safety Using SafetyBERT and Confidence-Based Active Learning

aakdani1@gmail.com
aakdani1@gmail.com

June 30, 2025

Abstract

This work addresses the challenges of extracting actionable safety insights from unstructured accident narratives in mining operations by leveraging advanced natural language processing techniques. It highlights the limitations of general-purpose language models when applied to domain-specific texts and presents the development and application of specialized transformer-based models, such as SafetyBERT, which are pretrained on mining safety corpora to capture nuanced terminology and contextual relationships. The integration of active learning frameworks employing confidence-based query selection optimizes annotation efficiency, reducing the manual labeling burden while enhancing model performance. Methodologies for data acquisition, preprocessing, and corpus construction are detailed, emphasizing the importance of linguistic diversity and the balance between practical narratives and academic content. The study further explores model evaluation metrics, iterative training with human-in-the-loop feedback, and error analysis to ensure robustness and interpretability. Applications demonstrate the deployment of these systems in mining operations and their transferability to other high-risk domains, including construction, healthcare, and transportation safety. The findings underscore the potential for domain-adapted language models combined with intelligent annotation strategies to transform safety management by enabling proactive hazard mitigation and scalable, data-driven decision-making in complex industrial environments.

1 Introduction

The analysis of accident data in mining safety is complicated by the unstructured and heterogeneous nature of incident narratives, which often contain nuanced contextual information not captured by structured metadata. Traditional manual review of such data is not only labor-intensive but also limited in its ability to systematically uncover latent patterns and relationships that are critical for effective risk mitigation and safety management. The integration of advanced natural language processing (NLP) techniques, particularly those leveraging domain-adapted language models, has emerged as a transformative approach for extracting actionable knowledge from these vast repositories of accident reports [1]. Recent advancements in NLP, especially the development of transformer-based models such as BERT and its domain-specific variants, have enabled more sophisticated analysis of free-text accident narratives [2][3]. These models are capable of capturing complex interdependencies between safety factors, which is essential for understanding the multifaceted causes of incidents in high-risk environments like mining. However, the direct application of general-purpose language models to specialized domains often results in suboptimal performance due to the mismatch between the model's pretraining data and the unique linguistic characteristics of domain-specific texts [2][3]. This challenge is particularly pronounced in occupational safety, where the terminology, context, and reporting style differ significantly from general language corpora [2][3]. To address these limitations, researchers have developed domain-adapted models such as SafetyBERT, which are pretrained on large corpora of safety-related documents to better capture the semantics and context of accident reports. Danish et al. demonstrate that such domain-specialized architectures not only improve the accuracy of information extraction but also enhance the interpretability of the results, supporting more balanced and reliable categorization of safety incidents. The adoption of these models facilitates scalable

and resource-efficient deployment in occupational safety management frameworks, laying the groundwork for more proactive and data-driven safety interventions. Despite these advances, the creation of high-quality labeled datasets for training and fine-tuning domain-specific models remains a significant bottleneck. Manual annotation of accident narratives is both time-consuming and costly, especially given the volume and complexity of the data. To mitigate this challenge, active learning strategies have been proposed, wherein the model iteratively selects the most informative or uncertain samples for annotation, thereby maximizing the efficiency of the labeling process. By employing confidence-based sampling, the framework ensures that annotation efforts are focused on examples that are most likely to improve model performance, reducing the overall annotation burden while accelerating the development of robust question-answering systems for mining safety [2]. The combination of NLP techniques such as semantic-rich embeddings, dimensionality reduction, and clustering further enhances the analytical depth of safety data analysis. These methods enable the discovery of deep insights into accident causation and prevention, surpassing the capabilities of traditional manual approaches. The systematic extraction and analysis of critical information from unstructured narratives not only inform policy and safety management but also support the development of preventive strategies tailored to the unique risks of mining operations. Moreover, the ongoing evolution of NLP, driven by advances in machine learning and deep learning, continues to expand the potential for automated safety analysis across diverse industrial domains [1][3]. The integration of specialized language models with intelligent sampling strategies represents a significant methodological advancement, offering scalable solutions for high-stakes environments where timely and accurate safety insights are paramount [2][3]. The adoption of such frameworks is poised to transform the landscape of occupational safety, enabling more effective risk mitigation and ultimately contributing to safer mining operations [1][2].

2 Background and Motivation

2.1 Significance of Accident Prevention in Mining and Related Industries

Accident prevention in mining and related industries is of paramount importance due to the inherently hazardous nature of these environments and the significant human, economic, and operational consequences associated with industrial incidents. Mining operations, in particular, are characterized by complex processes, heavy machinery, and challenging working conditions, all of which contribute to a heightened risk of accidents. The analysis of mining accident reports, which typically include both structured metadata and detailed unstructured narratives, is essential for uncovering the underlying causes, mechanisms, and outcomes of incidents. These reports serve as a foundational resource for both qualitative and quantitative investigations aimed at enhancing occupational safety and operational efficiency. The unstructured narratives found in accident documentation are especially valuable, as they provide context-rich descriptions that can reveal subtle contributing factors and causal chains not readily apparent in structured data alone. Extracting actionable knowledge from these narratives is crucial for developing effective preventive strategies. However, the complexity and volume of such data present significant analytical challenges, necessitating the adoption of advanced Natural Language Processing (NLP) techniques to systematically process and interpret the information. The integration of NLP with dimensionality reduction and clustering methods enables researchers to identify patterns and trends in accident causation, supporting the development of targeted interventions and policy recommendations. The significance of accident prevention extends beyond immediate safety concerns. Effective prevention strategies contribute to long-term improvements in workforce well-being, reduction of operational disruptions, and minimization of financial losses due to downtime, compensation claims, and regulatory penalties. Furthermore, the insights gained from accident analysis inform the design of safer work environments, the implementation of robust safety management systems, and the continuous improvement of industry standards. The ongoing evolution of NLP and machine learning technologies has further expanded the potential for extracting deep insights from accident data, surpassing the limitations of traditional manual analysis. By leveraging these advancements, safety researchers and practitioners are better equipped to address the multifaceted challenges of accident prevention in mining and related sectors. The systematic extraction and interpretation of knowledge from accident reports not only enhance the analytical depth of safety research but also provide a robust framework for risk mitigation and the proactive management of industrial hazards. In summary, the prevention of accidents in mining and related industries is a critical objective that underpins the sustainability

and resilience of these sectors. The strategic use of advanced analytical techniques, particularly those capable of handling unstructured textual data, is instrumental in transforming raw accident narratives into actionable safety insights. This transformation is essential for informing evidence-based decision-making, shaping effective safety policies, and ultimately safeguarding the lives and livelihoods of workers in high-risk industrial environments [1].

2.2 Impact of Accidents on Human, Economic, and Operational Factors

Accidents in industrial environments, particularly in mining, exert profound effects on human, economic, and operational dimensions. The human impact is often the most immediate and severe, manifesting as injuries, fatalities, and long-term health consequences for workers. These incidents not only disrupt the lives of affected individuals but also have cascading effects on their families and communities. The psychological burden on survivors and witnesses can be substantial, leading to decreased morale and productivity within the workforce [2][1]. The authors of [2] indicate that understanding the intricate relationships between safety factors and accident outcomes is essential for developing targeted interventions that mitigate these human costs. From an economic perspective, accidents impose significant direct and indirect costs on organizations and society. Direct costs include medical expenses, compensation payments, and legal liabilities, while indirect costs encompass lost productivity, equipment damage, and increased insurance premiums. The integration of advanced NLP techniques, such as text mining and classification, enables the systematic extraction of critical information from accident reports, which can inform strategies to reduce these economic burdens. By identifying patterns and risk factors embedded in unstructured data, organizations can prioritize preventive measures and allocate resources more efficiently, ultimately reducing the frequency and severity of costly incidents. Operationally, accidents disrupt the continuity of industrial processes, leading to downtime, delays, and reduced output. The complexity and unstructured nature of accident data present challenges for traditional analysis methods, often obscuring the root causes of operational disruptions. Leveraging NLP and machine learning frameworks allows for the extraction of actionable insights from vast repositories of accident narratives, supporting the development of robust safety management systems [1]. These systems can enhance decision-making by providing timely and relevant information, thereby improving risk mitigation and operational resilience. The multifaceted impact of accidents underscores the necessity for comprehensive analytical approaches that address human, economic, and operational factors simultaneously. The use of domain-adapted models and intelligent sampling strategies, as discussed in recent research, offers a scalable and resource-efficient pathway to enhance safety outcomes in high-risk industries [2]. By systematically analyzing accident data, organizations can not only safeguard human lives but also achieve substantial economic savings and maintain operational stability.

2.3 Challenges in Traditional Safety Data Analysis

Traditional safety data analysis in industrial domains such as mining faces significant challenges, particularly when dealing with unstructured accident reports and complex metadata. Historically, safety analysis has relied on manual review and categorization of incident narratives, which is both time-consuming and susceptible to human bias and inconsistency. The sheer volume and heterogeneity of data, including free-text narratives, categorical variables, and structured metadata, complicate the extraction of actionable insights [2][4]. Manual annotation of such data is resource-intensive, often requiring domain expertise to interpret nuanced descriptions of incidents, contributing to high costs and slow turnaround times [5][2]. Conventional clustering and classification techniques, such as k-means, have been widely used to group similar accident cases based on feature similarity. However, these methods struggle with high-dimensional, semantically rich text data, as they are not inherently designed to capture the contextual and linguistic subtleties present in narrative accident reports. The inability of traditional models to effectively process and interpret unstructured text limits their utility in uncovering latent patterns and trends that are critical for proactive safety management [1][4]. Furthermore, much of the valuable information embedded in accident narratives is not captured by structured metadata, leading to a loss of context and potentially overlooking key contributing factors to incidents [6]. Another challenge arises from the domain specificity of safety data. General-purpose language models and analytical frameworks often fail to account for the unique terminology, abbreviations, and context-dependent meanings prevalent in mining safety documentation [2][7]. This mismatch results in suboptimal performance when applying off-the-shelf models to specialized datasets, as these

models lack the necessary domain adaptation to accurately interpret and classify safety-related information [2][7]. The authors of [8] indicate that even advanced transformer-based models like BERT, while powerful in general NLP tasks, require further adaptation and fine-tuning to achieve satisfactory results in specialized fields such as occupational safety. The annotation bottleneck is further exacerbated by the need for large, high-quality labeled datasets to train and validate machine learning models. Manual labeling is not only labor-intensive but also prone to inconsistencies due to subjective interpretations of incident descriptions [5]. Active learning strategies have been proposed to mitigate this issue by prioritizing the annotation of the most informative or uncertain samples, thereby reducing the overall annotation burden while maintaining or improving model performance [9][5][10]. However, integrating such strategies into traditional workflows remains a challenge, particularly in environments where safety-critical decisions depend on timely and accurate analysis. Moreover, traditional safety data analysis methods often lack scalability and adaptability. As new types of incidents emerge and operational contexts evolve, static analytical frameworks struggle to keep pace with the dynamic nature of industrial environments [4]. This limitation hampers the ability to rapidly identify emerging hazards or shifts in risk patterns, which is essential for effective safety management. The complexity of accident causation, involving interdependencies among multiple safety factors, further complicates the modeling process, necessitating more sophisticated approaches capable of capturing these relationships [2]. In summary, the challenges in traditional safety data analysis stem from the unstructured and domain-specific nature of accident reports, the limitations of conventional analytical methods in handling complex text data, the high cost and subjectivity of manual annotation, and the lack of scalability in adapting to evolving safety landscapes [1][2][6]. Addressing these challenges requires the integration of advanced NLP techniques, domain-adapted models, and efficient data annotation strategies to enable more effective and scalable safety analysis in high-stakes industrial settings.

2.4 Role of Unstructured Safety Texts in Knowledge Discovery

Unstructured safety texts, such as accident narratives and incident reports, are a rich source of experiential knowledge that is often inaccessible through structured data fields alone. These narratives typically contain nuanced descriptions of the sequence of events, contributing factors, and environmental conditions surrounding safety incidents, which are not captured in predefined categorical fields. The inherent complexity and context-rich nature of these texts present both an opportunity and a challenge for knowledge discovery in industrial safety domains. The extraction of actionable insights from unstructured safety texts is essential for understanding the underlying causes of accidents and for informing the development of targeted safety interventions. Natural Language Processing (NLP) has emerged as a cornerstone methodology for systematically analyzing these texts, enabling researchers to identify patterns, trends, and relationships that would otherwise remain hidden. The application of advanced NLP techniques, such as tokenization, stopword removal, and contextual embedding generation, enhances the interpretability and analytical depth of unstructured data, facilitating the discovery of emergent risk factors and themes. Text mining, as a subset of NLP, allows for the efficient processing of large corpora of accident reports, supporting the identification of critical information that manual review would likely overlook. The integration of dimensionality reduction and clustering techniques further supports the management of high-dimensional data, enabling the extraction of deep insights into accident causation and prevention. The combination of these methods provides a robust framework for knowledge discovery, surpassing the limitations of traditional manual analysis and supporting the development of frameworks that inform policy, safety management, and preventive strategies. The role of unstructured safety texts extends beyond mere documentation; they serve as repositories of collective experience and tacit knowledge within high-risk industries such as mining, construction, and chemical processing [1]. Accurate extraction of knowledge from these texts is significant for improving the efficiency of safety management, as it enables the identification of experience-based knowledge that can inform the intelligent generation of safety measures. Machine learning approaches, particularly those that leverage semantic relationships within text data, have demonstrated the ability to intelligently mine knowledge from unstructured sources, overcoming the limitations of rule-based systems that depend heavily on the richness of predefined rule databases [11]. Recent advancements in large language models and domain-specific adaptations, such as SafetyBERT, have further expanded the potential for extracting critical safety insights from unstructured accident reports [12]. These models, when combined with intelligent sampling strategies and active learning frameworks, enable scalable and efficient knowledge discovery even in resource-limited scenarios [2][13]. The ongoing evolution of

NLP, driven by advances in machine learning and deep learning, continues to enhance the capacity for systematic knowledge extraction from unstructured safety texts, ultimately contributing to improved safety outcomes and risk mitigation across diverse industrial domains [1][12]. The integration of structured and unstructured data types allows for more holistic analyses, leading to richer insights and more targeted recommendations for safety interventions. By leveraging the strengths of NLP, dimensionality reduction, and clustering, researchers are equipped to extract actionable knowledge from vast repositories of accident narratives, supporting the development of effective safety analysis tools for high-stakes environments [1].

3 Literature Review

3.1 Occupational Safety in High-Risk Industries

3.1.1 Mining Industry Safety Challenges

The mining industry is characterized by a complex operational environment that inherently exposes workers and assets to significant safety risks. These risks stem from a combination of hazardous working conditions, the use of heavy machinery, and the unpredictable nature of geological formations. Accident data analysis in mining is particularly challenging due to the unstructured and heterogeneous nature of the data, which often consists of narrative accident reports, incident logs, and free-text safety observations. The inherent complexity of these data sources complicates the extraction of actionable insights, as traditional statistical methods are ill-suited for processing unstructured textual information. Natural Language Processing (NLP) has emerged as a transformative approach for addressing these challenges by enabling the systematic analysis of large volumes of accident narratives. The application of NLP techniques allows researchers to extract critical information regarding the causes, contributing factors, and outcomes of mining incidents, which are often embedded within detailed textual descriptions. The integration of advanced NLP methods, such as text classification and clustering, supports the identification of patterns and trends that may not be readily apparent through manual review. This analytical depth is essential for informing safety management strategies and developing targeted interventions aimed at risk mitigation. Mining accident reports serve as a foundational resource for both qualitative and quantitative safety analyses. These reports typically contain rich contextual information, including descriptions of hazardous events, environmental conditions, and human factors. The significance of mining accident reports lies in their ability to provide comprehensive insights into the mechanisms and consequences of incidents, thereby supporting the development of evidence-based safety policies and operational improvements [1]. However, the sheer volume and linguistic diversity of these reports present substantial obstacles to efficient analysis. For instance, accident narratives may vary widely in terminology, structure, and detail, further complicating efforts to standardize and interpret the data [2]. The challenges associated with mining industry safety are further exacerbated by the need for timely and accurate identification of emerging hazards. Manual review of accident reports is not only labor-intensive but also prone to inconsistencies and oversight, especially when dealing with large datasets. The adoption of automated text mining and NLP frameworks addresses these limitations by enabling scalable and reproducible analysis of safety data. By leveraging dimensionality reduction and clustering techniques, researchers can distill vast repositories of accident narratives into actionable knowledge, facilitating the discovery of latent risk factors and supporting proactive safety management [1]. Despite the promise of NLP-driven approaches, several methodological challenges persist. The development of domain-specific language models, such as SafetyBERT, is motivated by the recognition that general-purpose models often fail to capture the nuanced terminology and context-specific knowledge present in mining safety texts. The resource-intensive nature of manual data annotation further complicates model development, necessitating the exploration of active learning strategies to optimize the use of limited labeled data [2][5]. Active learning frameworks iteratively select the most informative samples for annotation, thereby enhancing model performance while minimizing annotation costs. The distribution of textual data in mining safety corpora also influences analytical outcomes. For example, academic abstracts, while comprising a smaller proportion of documents, contribute a disproportionately large share of tokens due to their linguistic density. In contrast, accident narratives, which make up the majority of documents, are characterized by greater variability in language and structure [2]. This heterogeneity underscores the importance of robust methodological design and the careful selection of analytical techniques to ensure the reliability and

generalizability of findings. The examination of mining accident narratives using advanced NLP and text mining techniques has demonstrated substantial utility in uncovering hidden risk factors and emergent safety themes. Content analysis supported by software tools enables the efficient processing of extensive datasets, allowing for the identification of trends and relationships that inform both policy and operational decision-making. The integration of these analytical frameworks into mining safety management systems holds significant potential for improving occupational safety outcomes and reducing the incidence of high-consequence events [1]. In summary, the mining industry faces multifaceted safety challenges rooted in the complexity of its operational environment and the unstructured nature of its safety data. The adoption of advanced NLP techniques, coupled with domain-specific modeling and active learning strategies, represents a promising pathway for enhancing the efficiency and effectiveness of safety analysis in this high-risk sector [2][1].

3.1.2 Construction and Manufacturing Safety Parallels

The intersection of construction and manufacturing safety research reveals significant methodological and practical parallels, particularly in the adoption of advanced natural language processing (NLP) and machine learning techniques for accident analysis and risk mitigation. Both industries are characterized by high-risk operational environments, where the consequences of safety lapses can be severe, necessitating robust systems for incident detection, analysis, and prevention. The abundance of unstructured textual data, such as accident reports and safety narratives, presents a shared challenge and opportunity for extracting actionable insights to inform safety interventions [1][4]. In both construction and manufacturing, NLP has emerged as a transformative tool for automating the interpretation of textual data, enabling the systematic extraction of critical information from large corpora of incident reports. The application of text mining and classification algorithms facilitates the identification of patterns, trends, and latent risk factors that may not be readily apparent through manual review [1]. For instance, the use of semantic and syntactic analysis in construction safety management has led to improved accuracy in identifying hazards and supporting worker well-being, a trend mirrored in manufacturing where similar techniques are leveraged to enhance health and safety management systems [4]. Clustering algorithms, particularly k-means, have been widely adopted in both sectors to categorize accident types, identify precursors, and support comprehensive risk assessment. The systematic grouping of accident attributes through unsupervised learning enables researchers to uncover meaningful patterns that inform targeted safety interventions. However, the effectiveness of these clustering approaches is highly contingent on the quality of data preprocessing and feature integration, underscoring the necessity for meticulous data preparation in both construction and manufacturing contexts. The integration of dimensionality reduction techniques, such as UMAP, t-SNE, and PCA, further enhances the interpretability of high-dimensional text data, supporting nuanced analyses that can inform policy development and safety management strategies. The scalability and adaptability of these analytical frameworks are crucial for their application in large-scale, safety-critical environments. Both industries benefit from frameworks that can efficiently process vast amounts of unstructured data, enabling real-time monitoring and proactive risk identification [1]. The transition from retrospective statistical analysis to proactive, data-driven safety management is facilitated by the ability of NLP-based systems to transform sparse incident records into actionable narratives, providing deeper insights into the interplay of diverse factors influencing accident outcomes [14]. A notable parallel lies in the challenge of domain adaptation for language models. General-purpose NLP models often struggle to capture the specialized vocabulary and context-specific nuances present in construction and manufacturing safety narratives. This limitation has driven the development of domain-specific models, such as SafetyBERT, which are fine-tuned on industry-specific corpora to enhance performance on classification and information extraction tasks [2]. The computational demands of training such models, coupled with the resource-intensive nature of manual data annotation, have prompted the adoption of active learning strategies. By employing intelligent, confidence-based sampling methods, these approaches minimize the requirement for labeled data while maintaining high model performance, a strategy that is equally relevant in both construction and manufacturing safety research [9][15]. The integration of metadata variables and semantic-rich embeddings with clustering and classification algorithms further supports the discovery of latent patterns in accident narratives, facilitating the development of targeted safety interventions and informing policy decisions [1]. The interpretability of these models is enhanced when combined with domain knowledge, enabling stakeholders to translate analytical findings into actionable safety measures. The convergence of methodological innovations in construction

and manufacturing safety research underscores the potential for cross-industry knowledge transfer. Techniques developed in one sector, such as the use of question-answering systems to extract experiential knowledge from safety hazard texts, can be adapted to the other, improving the efficiency of safety management and the generation of intelligent, context-specific interventions [11]. The empirical validation of these approaches across diverse industrial settings highlights the importance of methodological flexibility and the continuous refinement of analytical frameworks to address the evolving challenges of occupational safety in high-risk environments [1][4].

3.1.3 Historical Trends in Industrial Accident Prevention

The evolution of industrial accident prevention has been shaped by the interplay between technological progress, regulatory frameworks, and the increasing complexity of high-risk work environments. Early efforts in accident prevention were predominantly reactive, relying on manual investigation and reporting of incidents. These traditional approaches, while foundational, were limited by the subjective interpretation of events and the labor-intensive nature of data collection and analysis. As industries expanded and the volume of accident data grew, the need for more systematic and scalable methods became apparent. The introduction of digital record-keeping and the accumulation of large-scale accident databases marked a significant shift. However, much of this data remained unstructured, residing in narrative reports and free-text fields, which posed challenges for conventional analytical techniques. Manual review of such records was not only time-consuming but also prone to inconsistencies and oversight, especially when attempting to identify subtle patterns or emerging risk factors across diverse operational contexts [1][12]. The limitations of manual analysis underscored the necessity for automated tools capable of extracting actionable insights from complex datasets. Advancements in natural language processing (NLP) have catalyzed a methodological transformation in accident prevention research. The adoption of NLP techniques, including semantic-rich embeddings, dimensionality reduction, and clustering, has enabled the systematic extraction of knowledge from unstructured accident data, surpassing the constraints of earlier manual approaches. Text mining, a subset of NLP, has proven particularly effective in uncovering hidden patterns, trends, and relationships within large corpora of accident reports, thereby informing more proactive safety management strategies. The ability to process vast datasets efficiently has facilitated the discovery of risk factors and emergent themes that would likely remain undetected through traditional review methods [1][12]. The integration of machine learning and deep learning into safety research has further expanded the analytical capabilities available to practitioners. For instance, the development of hybrid systems that combine principal component analysis with neural networks and support vector machines has enabled more nuanced prediction of accident severity and classification of injury levels. These approaches have allowed for the modeling of complex, multi-class outcomes, reflecting the multifaceted nature of industrial accidents [16]. The evolution of pre-trained language models, such as BERT, has introduced bidirectional contextual understanding, which is critical for interpreting the nuanced language found in accident narratives [17]. Fine-tuning these models for domain-specific tasks has demonstrated substantial improvements in classification accuracy and the extraction of relevant safety information [2][17]. Despite these technological advances, the application of general-purpose models to specialized domains like mining safety has revealed persistent challenges. Domain adaptation is essential, as generic models often fail to capture the unique terminology and contextual relationships inherent in high-risk industries. The development of domain-specialized models, such as safetyBERT and safetyALBERT, represents a significant step forward. These models, pre-trained on extensive corpora of safety documents and academic literature, have shown superior performance in occupational safety applications, effectively capturing interdependencies between safety factors and supporting more balanced, category-wise analysis [2]. Another historical trend in accident prevention is the increasing emphasis on scalable and resource-efficient strategies for model improvement. Manual data annotation, while valuable, is resource-intensive and often impractical for large datasets. Active learning approaches, which prioritize the annotation of data points where model confidence is low, have emerged as a solution to this bottleneck. By intelligently sampling the most informative examples, active learning reduces the annotation burden while accelerating model performance gains, making it particularly suitable for high-stakes industrial environments where rapid adaptation is crucial [13][12]. The role of human expertise remains integral in the deployment and evaluation of automated safety analysis tools. Subject matter experts are essential for assessing model vocabulary, validating outputs, and ensuring that the system's recommendations align with operational realities. The structure and content of queries posed

to extractive models must be carefully considered to maximize the relevance and accuracy of the information retrieved [18][12]. This interplay between automated analysis and expert oversight reflects a broader trend toward hybrid systems that leverage both computational power and domain knowledge. Recent developments have also highlighted the importance of extracting experiential knowledge from safety hazard texts to inform management measures and improve overall safety efficiency. The creation of intelligent question-answering systems capable of synthesizing experience-based insights from unstructured data exemplifies the ongoing shift toward proactive, knowledge-driven accident prevention [11]. Collectively, these historical trends illustrate a trajectory from manual, reactive approaches toward automated, data-driven, and domain-adapted methodologies in industrial accident prevention. The continuous refinement of NLP and machine learning techniques, coupled with active learning and expert validation, is shaping a new era of occupational safety in high-risk industries [1][2].

3.2 Natural Language Processing for Safety Knowledge Extraction

3.2.1 Traditional Information Extraction Approaches

Traditional information extraction (IE) approaches in the context of safety knowledge mining have historically relied on a combination of manual rule-based systems, shallow machine learning models, and structured data extraction techniques. These methods were primarily designed to process structured or semi-structured data, often struggling with the complexity and variability inherent in unstructured accident reports and incident narratives. The extraction of meaningful safety insights from such unstructured textual data has been a longstanding challenge, particularly due to the diversity of language, domain-specific terminology, and the nuanced context in which safety events are described [2]. Early IE systems typically employed handcrafted rules and pattern-matching algorithms to identify relevant entities, relationships, and events within text. These systems required significant domain expertise to encode linguistic patterns and were highly sensitive to variations in language use. As a result, their scalability and adaptability to new domains or evolving safety standards were limited. For example, extracting accident causes or risk factors from narrative reports often necessitated the manual definition of keyword lists and syntactic templates, which could not easily generalize across different industrial sectors or reporting styles [11][4]. The advent of statistical machine learning introduced more flexible approaches, such as supervised classifiers and sequence labeling models, including support vector machines and conditional random fields. These models leveraged annotated corpora to learn patterns associated with safety-related entities and events. However, their effectiveness was constrained by the availability and quality of labeled training data, which is particularly scarce in specialized domains like mining safety. The annotation process itself is resource-intensive, requiring expert knowledge to accurately label complex safety concepts and relationships [19][2]. Slot identification and intent understanding represent two foundational tasks in traditional IE pipelines. Slot identification focuses on recognizing and categorizing specific entities, such as names, dates, locations, or technical terms, within a query or document. Traditionally, separate models were trained for each task, necessitating the development and maintenance of multiple specialized systems. This compartmentalized approach often led to inefficiencies and limited the ability to capture the full semantic context of safety narratives [20]. Text mining techniques, including topic modeling and clustering, have also been applied to accident reports to uncover latent themes and patterns. For instance, the integration of deep learning with latent Dirichlet allocation (LDA) models has enabled the automatic extraction and visualization of accident information, supporting risk factor identification and safety management decisions. These methods facilitate the construction of risk networks by linking extracted knowledge from diverse textual sources, yet they still depend on the quality of initial feature engineering and the representativeness of the input data [11][4]. Despite these advancements, traditional IE approaches often fall short when confronted with the intricacies of unstructured safety data. The heterogeneity of reporting formats, the prevalence of domain-specific jargon, and the implicit nature of causal relationships in accident narratives pose significant obstacles. Natural language processing (NLP) offers a suite of tools to address these challenges, but the transition from rule-based and shallow learning methods to more sophisticated models has been gradual. The potential of unstructured incident narratives for enhancing occupational safety remains underexplored, largely due to the limitations of earlier extraction techniques. Recent efforts have focused on developing systematic frameworks for collecting and processing safety-related textual data from multiple sources, including academic literature and regulatory reports. These frameworks employ a combination of API-based extraction systems and multi-method

approaches tailored to the specific characteristics of each data source. Such strategies aim to overcome the fragmentation of safety knowledge and enable more comprehensive analysis by integrating information from diverse domains [2]. The limitations of traditional IE approaches underscore the need for more advanced, domain-adapted NLP techniques capable of handling the complexity and scale of unstructured safety data. The emergence of transformer-based language models and active learning frameworks represents a significant step forward, offering the potential to automate and enhance the extraction of critical safety insights with greater accuracy and efficiency [21][19].

3.2.2 Evolution of Deep Learning in Text Mining

The evolution of deep learning in text mining has fundamentally transformed the landscape of natural language processing (NLP), enabling the extraction of complex knowledge from unstructured textual data. Early approaches in text mining relied heavily on rule-based systems and shallow machine learning models, which often struggled to capture the nuanced semantics and contextual dependencies inherent in natural language. The advent of deep learning, particularly with the introduction of transformer-based architectures, marked a significant leap forward in the ability to model language at scale and with greater fidelity. BERT (Bidirectional Encoder Representations from Transformers) exemplifies this shift, as it leverages bidirectional context to enhance the understanding of textual information. Unlike earlier unidirectional models such as GPT, which process text sequentially and thus have limited context awareness, BERT processes input in both directions, allowing for a richer representation of meaning. This bidirectional approach has proven especially effective for tasks that require deep comprehension, such as question answering and information extraction [22]. The authors of [23] indicate that integrating BERT with information retrieval systems, as in the BERTserini framework, yields substantial improvements in open-domain question answering, highlighting the practical impact of deep learning advancements. Despite these advances, the application of general-purpose models like BERT to specialized domains, such as mining safety, presents unique challenges. Domain-specific texts often contain specialized terminology and context that are not well represented in the corpora used for pre-training general models. As a result, the performance of these models can degrade when applied to domain-specific tasks [8][24]. According to [8], further pre-training or fine-tuning on domain-specific datasets is necessary to bridge the gap between open-domain and specialized data distributions. This process allows the model to internalize domain-relevant vocabulary and patterns, thereby improving its effectiveness in extracting critical knowledge from technical documents. The need for domain adaptation is particularly pronounced in fields like finance and biomedicine, where the language used diverges significantly from general corpora. Yuan [24] states that while fine-tuning pre-trained BERT models can achieve state-of-the-art results with relatively few labeled examples, the unique vocabulary and structure of financial texts necessitate additional adaptation steps. Similarly, Yugu et al. **0210406** outline that comprehensive benchmarks tailored to biomedical NLP have accelerated progress by enabling direct comparisons among specialized language models, underscoring the importance of domain-specific evaluation and training. Another significant development in deep learning for text mining is the emergence of active learning strategies. Manual annotation of large datasets is resource-intensive, particularly in high-stakes domains where expert knowledge is required. Active learning addresses this challenge by intelligently selecting the most informative samples for annotation, thereby reducing the overall labeling effort while maintaining or even improving model performance [25][26]. Sumanth Prabhu et al. [25] demonstrate that active learning with BERT can achieve comparable F1-scores to fully supervised approaches while reducing labeling costs by up to 85%. This efficiency is crucial for scaling NLP solutions in domains where annotated data is scarce or expensive to obtain. The integration of knowledge distillation with active learning further enhances the efficiency of model training. Boreshban et al. [9] describe a framework where a large, fine-tuned teacher model (such as BERT) transfers its knowledge to a smaller student model through distillation, guided by active learning-based sample selection. This approach not only reduces computational requirements but also maintains high performance, making it suitable for deployment in resource-constrained environments. Model size and computational efficiency have also become central considerations in the evolution of deep learning for text mining. PankajKumar et al. [7] highlight that smaller models like PureMechBERT can process samples significantly faster than larger counterparts, such as DeBERTa LARGE v3, without substantial loss in precision. This balance between speed and accuracy is particularly valuable for real-time information extraction systems, where operational efficiency and energy sustainability are critical. Extractive and abstractive methods represent two complementary strategies

in deep learning-based text mining. While abstractive systems aim to synthesize new answers by generating text, extractive approaches select relevant spans directly from source documents. Frermann [27] notes that extractive systems, though conceptually simpler, can serve as effective first steps toward more sophisticated answer generation, especially when leveraging the strengths of deep learning models in identifying salient information. The evolution of deep learning in text mining has also facilitated the development of interpretable frameworks for complex applications, such as traffic crash analysis. Hao Zhen and Jidong J. Yang [14] present a large language model-centered framework that provides contextual and interpretable insights from accident reports, supporting targeted safety interventions. However, they acknowledge limitations in capturing multimodal information, suggesting that future advancements may involve integrating textual and non-textual data sources. Commonsense knowledge mining from pretrained models has further enriched the capabilities of deep learning systems, enabling them to reason beyond explicit textual content [28]. By incorporating external linguistic knowledge during pre-training, models can achieve more robust understanding and generalization, which is particularly beneficial for tasks involving nuanced or implicit information. In summary, the trajectory of deep learning in text mining reflects a continuous interplay between model innovation, domain adaptation, annotation efficiency, and operational scalability. The convergence of these advancements has enabled the extraction of actionable knowledge from unstructured text, laying the groundwork for sophisticated, domain-specific NLP applications in safety-critical industries [2][28][25][8].

3.2.3 Recent Advances in Transformer-Based Models

Recent advances in transformer-based models have significantly influenced the field of natural language processing, particularly in specialized applications such as safety knowledge extraction from unstructured accident reports. The introduction of models like BERT has marked a shift from traditional rule-based and early machine learning approaches to architectures capable of capturing complex semantic and syntactic relationships within text [18]. These models leverage self-attention mechanisms to encode contextual information, enabling a deeper understanding of language structure and meaning. One of the key strengths of transformer-based models is their ability to learn rich representations of language, which has been demonstrated through their capacity to encode subject-verb agreement and semantic roles, as well as to extract syntactic structures such as dependency and constituency trees [28]. This syntactic awareness is crucial for accurately interpreting the nuanced language often found in accident reports, where domain-specific terminology and complex sentence constructions are prevalent. The development of domain-adapted variants, such as BioBERT and BlueBERT in the biomedical field, illustrates the effectiveness of further pre-training general-purpose transformers on specialized corpora [3][29]. These models initialize with weights from general-domain BERT and are subsequently exposed to large-scale, domain-specific texts, allowing them to internalize the unique vocabulary and semantic patterns of the target field. This approach has been shown to outperform models trained solely on general data, particularly in tasks requiring deep domain knowledge [3]. In the financial sector, for example, FinBERT was created by further pre-training BERT on a financial corpus, resulting in superior performance on sentiment classification tasks compared to models lacking this additional adaptation [24]. Despite their impressive capabilities, transformer-based models present notable challenges when applied to specialized domains. The high dimensionality of sentence embeddings generated by advanced models such as SBERT, often spanning hundreds or thousands of dimensions, enables the capture of subtle semantic relationships between terms, even when their surface forms differ significantly. However, this complexity introduces computational burdens, making downstream tasks like clustering and pattern recognition more demanding in terms of both processing power and interpretability [1]. Additionally, the large scale of models such as BERT-base entails substantial training times and hardware requirements, which can be prohibitive in industrial settings where rapid deployment and scalability are essential [8][30]. To address these challenges, recent research has explored strategies such as mixed-domain pre-training, where models are first trained on general corpora and then further adapted to the target domain, as well as the use of pre-trained transformers as auxiliary memory modules in neural machine translation and other tasks [3]. These approaches have demonstrated that leveraging pre-trained knowledge, either as initialization or as an external resource, can enhance model performance and efficiency, particularly when labeled data is scarce or costly to obtain [28]. In the context of safety knowledge extraction, the integration of transformer-based models with intelligent sampling strategies, such as active learning, offers a promising avenue for reducing annotation costs while maintaining high model accuracy. By prioritizing the labeling of data points

where the model exhibits low confidence, active learning frameworks can efficiently guide the annotation process, ensuring that the most informative examples are selected for human review [31]. This is especially relevant in high-stakes environments like mining safety, where the timely and accurate extraction of critical insights from unstructured reports can have direct implications for operational risk management. The evolution of transformer-based models has also facilitated the development of advanced question-answering systems tailored to specific domains. Early Q&A frameworks were predominantly rule-based, but the advent of deep learning and transformers has enabled the creation of models capable of understanding and reasoning over complex, heterogeneous data sources [18]. For instance, CrashLLM leverages text reasoning to analyze traffic crash data, outperforming traditional machine learning models and supporting data-driven decision-making for infrastructure improvements [16]. In financial applications, fine-tuned transformer models such as RoBERTa have been combined with techniques like stride shift and dense passage retrieval to handle long-form answers and prevent duplicate responses, further illustrating the adaptability of these architectures to domain-specific requirements [31]. The ongoing refinement of transformer-based models, including the development of specialized variants and the incorporation of efficient training and inference strategies, continues to expand their applicability in safety-critical domains. As computational resources and annotation budgets remain limiting factors, the combination of domain adaptation, active learning, and scalable model architectures represents a promising direction for future research and deployment in industrial safety analysis [8][30].

3.3 Question Answering Systems in Industrial Domains

3.3.1 General-Domain QA Systems

General-domain question answering (QA) systems have undergone significant evolution, primarily driven by advances in natural language processing (NLP) and the availability of large-scale, unstructured text corpora. These systems are typically trained on vast datasets such as Wikipedia, web text, and book corpora, which provide a broad linguistic foundation but often lack the specialized terminology and contextual nuances required for domain-specific applications. The foundational approach involves pretraining language models from scratch on these general corpora, enabling the models to capture a wide range of syntactic and semantic patterns. However, this generalization comes at the cost of limited effectiveness when applied to specialized fields, where domain-specific concepts and jargon are prevalent and often absent from general-domain data [3][32]. The architecture of modern general-domain QA systems is heavily influenced by transformer-based models, with BERT and its derivatives being particularly prominent. BERT leverages deep bidirectional context through Masked Language Modeling (MLM) and Next Sentence Prediction (NSP), allowing it to understand complex relationships within and between sentences [2]. This bidirectional encoding is crucial for tasks that require nuanced comprehension, such as extracting answers from unstructured text. The success of BERT has led to the development of subsequent models that further refine its capabilities, often by incorporating larger datasets or more sophisticated pretraining objectives [32]. These models have demonstrated strong performance on benchmark QA datasets, especially in open-domain settings where the diversity of language and topics is high [2][32]. Despite their strengths, general-domain QA systems face notable challenges when deployed in industrial or safety-critical environments. One major limitation is their reliance on annotated data for supervised learning. The process of labeling large volumes of training data is resource-intensive and often requires domain expertise, which is not always readily available [5]. This bottleneck restricts the scalability and adaptability of general-domain models to new or evolving domains. Furthermore, the generalization achieved by these models does not always translate to high accuracy in specialized contexts, where subtle distinctions and rare terminology can significantly impact the relevance and correctness of extracted answers [3][5]. To address the issue of answer selection and relevance, general-domain QA systems often employ multi-stage retrieval and ranking frameworks. An initial retrieval step, such as BM25, is used to identify candidate passages that are likely to contain the answer [24][33]. These candidates are then re-ranked using more sophisticated neural architectures, such as Siamese networks or Compare-Aggregate frameworks, to select the most contextually appropriate answer [24]. The integration of these retrieval and re-ranking mechanisms enhances the precision of QA systems, particularly in scenarios where the answer space is large and diverse. Khattab et al. [33] demonstrate that iterative refinement of retrieval strategies, even with weak supervision, can yield substantial improvements in answer quality over traditional bag-of-words

approaches. The adaptability of general-domain QA systems has also been explored in various industrial applications, including finance and construction safety. In the financial sector, for example, transformer-based models like BERT have been adapted to handle non-factoid answer selection, where the goal is to retrieve and rank passage-level texts that best address complex, context-dependent queries [24]. Similarly, in construction safety, the integration of deep learning models with domain knowledge has been proposed to enhance the efficiency and accuracy of QA systems, supporting the intelligent transformation of industry practices. These adaptations underscore the potential of general-domain architectures as foundational tools, but also highlight the necessity for domain-specific customization to achieve optimal performance [32][24]. The scalability and objectivity of general-domain QA systems make them attractive for large-scale data analysis, especially in contexts where manual review is impractical. Techniques such as semantic enhancement, contextualization through SBERT-based embeddings, and dimensionality reduction have been employed to improve clustering and pattern discovery in unstructured safety data [1]. However, the effectiveness of these methods is contingent on the ability of the underlying language models to accurately represent domain-specific semantics, which is often limited in general-domain pretrained models [3][1]. In summary, general-domain QA systems provide a robust baseline for automated information extraction and answer retrieval across a wide range of topics. Their reliance on large, diverse corpora and advanced neural architectures enables strong performance in open-domain settings. Nevertheless, their application in specialized industrial domains is constrained by challenges related to domain adaptation, data annotation, and the representation of specialized knowledge. These limitations have motivated the development of domain-specific models and active learning strategies, which aim to bridge the gap between general linguistic competence and the nuanced requirements of high-stakes industrial environments [3][32][5].

3.3.2 Domain Adaptation for Specialized QA

Domain adaptation for specialized question answering (QA) is a critical challenge when deploying advanced natural language processing (NLP) systems in industrial contexts such as mining safety. General-purpose language models, while powerful, often struggle to capture the nuanced terminology, reporting conventions, and implicit knowledge embedded in domain-specific corpora. This mismatch can result in suboptimal extraction of actionable insights from unstructured accident reports, where the stakes for accuracy and interpretability are high [2][14]. The process of adapting language models to specialized domains typically involves fine-tuning pre-trained architectures on curated datasets that reflect the target domain's linguistic and semantic characteristics. For instance, the development of SafetyBERT and safetyALBERT demonstrates that domain-specific pretraining and continued adaptation can yield models that are both robust and computationally efficient for occupational safety applications. Notably, safetyALBERT achieves strong performance despite a reduced parameter count, underscoring the value of parameter-efficient architectures in resource-constrained industrial settings. This efficiency is particularly relevant given the often limited availability of annotated data in specialized fields, where manual labeling is both time-consuming and costly. To address the scarcity of labeled data, active learning strategies have been proposed as a means to maximize annotation efficiency. By employing intelligent, confidence-based sampling, the model can prioritize the most informative or uncertain instances for human annotation, thereby accelerating the learning curve with minimal labeling effort [2]. This approach is especially advantageous in high-stakes environments, where rapid adaptation to evolving safety concerns is essential. Another key aspect of domain adaptation is the integration of context-aware data augmentation techniques. For example, leveraging large language models such as LLaMA3-8B, it is possible to generate semantically coherent and factually consistent augmentations of crash reports. This process involves applying linguistic transformations under preservation constraints, guided by expert prompts that ensure the augmented narratives remain faithful to the original safety context. Such augmentation not only enriches the training data but also enhances the model's ability to generalize across diverse incident scenarios. Explainability remains a central concern in safety-critical QA systems. The authors of [14] indicate that current efforts often overlook the need for transparent reasoning, which is vital for trust and accountability in industrial decision-making. Incorporating domain knowledge and systematic data transformation into the adaptation pipeline can help bridge this gap, enabling models to provide interpretable justifications for their outputs. Comparative studies in other specialized domains, such as political text analysis and materials science, further highlight the importance of tailored adaptation strategies. For instance, multi-task learning and knowledge distillation have been explored to adapt GPT models for political

corpora, while information extraction pipelines have been successfully learned by language models in materials informatics [7][17]. These findings suggest that the principles of domain adaptation, targeted fine-tuning, data augmentation, and active learning, are broadly applicable across diverse industrial QA tasks. The effectiveness of domain adaptation is also influenced by the choice of base model and the scale of pretraining data. Increasing the diversity and volume of pretraining corpora has been shown to improve downstream performance, although access to large, high-quality datasets remains a practical limitation in many specialized fields [34]. Furthermore, the use of Siamese networks for fine-tuning sentence representations, as implemented in SEDAN, demonstrates the utility of architectures that explicitly model contextual similarity between questions and candidate answers, thereby enhancing semantic understanding in domain-specific QA [19]. In operational QA systems, models such as BERTserini and bert-swiftqa exemplify the integration of retrieval and reader components, where the reader is fine-tuned to extract answer spans from domain-relevant contexts [23][18]. The performance of these systems is further validated by benchmarks such as SQuAD, where models like MiniLM and RoBERTa have demonstrated high precision and efficiency in answer extraction, suggesting that careful model selection and adaptation can yield significant gains even in specialized applications [35]. Overall, the literature underscores that successful domain adaptation for specialized QA hinges on a synergistic combination of domain-specific pretraining, active learning, context-aware augmentation, and explainability mechanisms. These strategies collectively enable the development of scalable, interpretable, and effective QA systems tailored to the unique demands of industrial environments such as mining safety [14][2][19].

3.3.3 Existing QA Systems in Safety and Healthcare

Existing question answering (QA) systems in safety and healthcare have evolved significantly with the advent of advanced natural language processing (NLP) models, particularly those based on transformer architectures. In safety-critical domains, the extraction of actionable knowledge from unstructured textual data, such as accident reports and incident narratives, is essential for effective risk management and prevention strategies. The integration of deep learning models, especially those pre-trained on large corpora and subsequently fine-tuned for domain-specific tasks, has enabled substantial improvements in the accuracy and efficiency of QA systems [32]. In the healthcare sector, domain-adapted models such as BioELECTRA have demonstrated strong performance across a variety of biomedical text mining tasks, including named entity recognition, relation extraction, and question answering. These models leverage large-scale biomedical datasets, such as PubMed abstracts and clinical notes, to capture the nuanced language and terminology unique to the medical field. The results from benchmarking studies indicate that fine-tuned transformer models can outperform traditional approaches in extracting relevant information and supporting clinical decision-making [3]. Similarly, in safety management, the application of BERT and its derivatives has shown promise in automating the retrieval of safety knowledge, thereby enhancing the efficiency of safety management processes on construction sites [32]. The use of large language models (LLMs) in safety and healthcare QA systems is not without challenges. One major limitation is the requirement for substantial amounts of high-quality, domain-specific data to achieve optimal performance. In many safety-critical industries, such data may be scarce, leading to models that provide only generic responses rather than detailed, context-specific insights. Furthermore, there is a lack of standardized metrics for evaluating the accuracy and robustness of LLM outputs in these domains, which complicates the assessment of system reliability. Security concerns, such as unauthorized data disclosure and vulnerability to adversarial attacks, also pose significant risks when deploying LLM-based QA systems in sensitive environments [36]. Recent research in transportation safety has explored the use of LLMs for analyzing accident reports and extracting structured information from unstructured narratives. For example, LLMs have been employed to classify crash severity and provide interpretable explanations that align with domain knowledge, supporting data-driven decision-making in urban transit systems. These applications underscore the potential of LLMs to enhance resource allocation and improve passenger safety, provided that careful engineering and fine-tuning are undertaken to address domain-specific requirements [14]. Maroa Mumtarin et al. [12] highlight that AI-driven systems can classify crash narratives and identify contributing factors, which is instrumental in developing targeted interventions for road safety. In industrial safety, unsupervised machine learning techniques, including clustering and text mining, have been effective in analyzing large-scale accident datasets. These approaches facilitate the identification of trends and associations that may not be apparent through manual analysis, enabling the discovery of hazardous

practices and the development of preventive measures. The automation of incident categorization and the extraction of semantic relationships from narrative data represent significant advancements over traditional manual review processes. The integration of clustering results with descriptive modeling allows for the clear categorization of risk factors and accident types, providing a structured basis for monitoring trends and informing policy decisions. Moreover, clustering analysis enables the classification of safety risk factors into direct causes and management deficiencies, supporting the prioritization of interventions and the tailoring of safety programs to address the most pressing hazards [1]. Despite these advancements, the deployment of multiple large models in specialized domains can impede scalability and necessitate frequent retraining as new data emerges. The performance of domain-adapted QA systems often depends on the quality of synthetic data used for self-supervised domain adaptation, which can be particularly challenging in niche fields with limited annotated resources [19]. The adoption of active learning strategies, where models iteratively select the most informative samples for annotation, has been proposed as a means to efficiently improve model performance with minimal labeled data, thereby addressing the resource-intensive nature of manual data annotation. Overall, existing QA systems in safety and healthcare are increasingly leveraging advanced NLP techniques and domain-specific adaptations to extract critical insights from unstructured data. While these systems have demonstrated substantial benefits in terms of efficiency, accuracy, and scalability, ongoing challenges related to data availability, evaluation standards, and security must be addressed to fully realize their potential in high-stakes industrial and healthcare environments [36][14][12][32].

3.4 Active Learning in Data-Scarce Environments

3.4.1 Fundamental Concepts and Strategies

Active learning has emerged as a crucial strategy for addressing the challenges posed by data scarcity in specialized domains such as mining safety, where annotated datasets are limited and manual labeling is resource-intensive. At its core, active learning is an iterative process in which a model selectively queries the most informative or uncertain data points for annotation, thereby maximizing learning efficiency while minimizing the annotation burden [5][26]. This approach is particularly advantageous in high-stakes environments, where the cost of errors is significant and the availability of domain experts for labeling is constrained. A fundamental concept underpinning active learning is the use of uncertainty-based sampling strategies. These strategies prioritize data instances for which the model exhibits the least confidence in its predictions, as these are expected to yield the greatest improvement in model performance upon annotation. The least confidence query strategy, for example, identifies samples where the predicted probability for the most likely class is lowest, signaling high uncertainty. By focusing annotation efforts on such samples, the model can rapidly refine its decision boundaries, especially in regions of the feature space that are underrepresented or ambiguous [5][26]. The integration of pre-trained language models, particularly those tailored to domain-specific contexts, further enhances the effectiveness of active learning frameworks. BERT-based architectures have demonstrated robust performance across a variety of NLP tasks, and their adaptability through fine-tuning makes them well-suited for specialized applications [17][37]. The emergence of domain-specific variants, such as SafetyBERT, exemplifies the trend of leveraging additional pretraining on relevant corpora to capture the nuanced language and semantics inherent in safety-related documents [2][37]. These specialized models are better equipped to extract critical insights from unstructured accident reports, outperforming general-purpose models that may lack sensitivity to domain-specific terminology and context [2][37]. Another key strategy involves the use of model distillation to balance computational efficiency with annotation effectiveness. Distillation enables the deployment of smaller, distilled versions of pre-trained models as proxies for querying, thereby reducing the computational overhead associated with active learning cycles without incurring substantial performance loss [26]. This approach is particularly beneficial when scaling active learning to large datasets or when computational resources are limited. The process of preparing data for active learning in safety-critical domains often requires sophisticated preprocessing pipelines. For instance, unstructured clinical or accident narratives must be segmented and labeled according to relevant frameworks, such as the SOAP (Subjective, Objective, Assessment, Plan) structure in clinical text, to provide meaningful seed data for model training [5]. The use of robust annotation frameworks ensures that the active learning model receives high-quality, contextually rich input, which is essential for effective learning in data-scarce settings. Dimensionality reduction techniques, such as UMAP, are frequently employed to manage

the high-dimensional embeddings generated by advanced NLP models like SBERT. These techniques facilitate the visualization and analysis of complex semantic spaces, enabling practitioners to better understand the distribution of uncertainty and the coverage of annotated samples [1]. By mapping dense vector representations to lower-dimensional spaces, researchers can identify clusters of similar cases and strategically target underrepresented regions for annotation. The synergy between intelligent sampling strategies, domain-adapted language models, and efficient computational techniques forms the foundation of scalable, effective active learning systems for mining safety and related fields. The iterative refinement of models through targeted annotation not only accelerates the development of high-performance question-answering systems but also ensures that these systems remain adaptable to evolving safety challenges and regulatory requirements [11][14]. The authors of indicate that integrating deep learning with hierarchical matching networks further enhances the precision and speed of information retrieval, which is critical for timely decision-making in industrial environments. By combining these fundamental concepts and strategies, active learning frameworks can transform the landscape of safety analysis, enabling the extraction of actionable insights from unstructured data with minimal manual intervention and maximal impact on operational safety [11][2][5][26].

3.4.2 Active Learning in NLP Applications

Active learning has emerged as a transformative strategy in natural language processing (NLP), particularly when addressing the challenges posed by data-scarce environments. In many specialized domains, such as mining safety, the availability of large, high-quality labeled datasets is limited due to the resource-intensive nature of manual annotation. This scarcity necessitates efficient approaches that maximize model performance while minimizing annotation costs. Active learning addresses this by iteratively selecting the most informative data points for labeling, thereby optimizing the annotation process and accelerating model improvement. A core principle of active learning in NLP is the use of uncertainty-based sampling strategies. These strategies prioritize instances for annotation where the model exhibits the least confidence in its predictions. By focusing on ambiguous or challenging examples, the model can rapidly learn from its mistakes and generalize better to unseen data. For instance, in the context of annotating unstructured clinical text, a robust active learning framework can leverage least-confidence query strategies to guide the selection of sentences for manual labeling, which in turn enhances the performance of BERT-based multiclass annotation models. This approach is not only applicable to clinical data but is also highly relevant for mining safety, where incident reports often contain complex, domain-specific language and nuanced contextual information. The integration of active learning with advanced language models, such as BERT and its domain-specific variants, has demonstrated significant improvements in annotation efficiency and model accuracy. The authors of [5] outline that by designing preprocessing algorithms for unstructured text and employing active learning as an integral component, it is possible to create scalable annotation pipelines. These pipelines can be reused across different datasets and tasks, further reducing the manual effort required for data labeling. Moreover, the active learning paradigm supports the development of specialized models that are better attuned to the linguistic and contextual peculiarities of high-stakes industrial environments. Despite these advantages, several challenges persist. The process of identifying the most informative samples for annotation is inherently dependent on the model's current state and may be influenced by biases in the initial seed data. Additionally, the effectiveness of active learning is closely tied to the quality of the underlying language model and the representativeness of the labeled data. In domains where narrative data is prevalent, such as accident reports in mining, the ambiguity and complexity of incident descriptions can complicate the active learning process. Clustering analysis and embedding-based representations can help mitigate some of these challenges by capturing semantic relationships and contextual cues that are not readily accessible through traditional keyword-based methods [1]. Furthermore, the adoption of active learning in NLP applications is often complemented by the use of pre-trained models and transfer learning. For example, leveraging pre-trained BERT models as a foundation allows for rapid adaptation to new tasks with minimal labeled data, especially when combined with active learning strategies [38][18]. This synergy enables the development of robust question-answering systems and annotation frameworks that are both efficient and scalable, even in data-constrained settings. In summary, active learning represents a critical advancement for NLP applications in data-scarce environments. By intelligently selecting data for annotation and integrating with powerful language models, it is possible to build effective, domain-specific tools for extracting actionable insights from unstructured text. This approach is particularly valuable in specialized fields

like mining safety, where the stakes are high and the cost of manual annotation is prohibitive [5][1][38].

3.4.3 Challenges in Annotating Safety Texts

Annotating safety texts for the development of domain-specific question-answering systems presents a range of unique challenges, particularly in data-scarce environments where active learning strategies are employed. One of the primary obstacles is the scarcity of sufficiently annotated data, which is essential for training deep learning models to accurately classify and extract relevant information from complex safety narratives. The lack of large, labeled datasets often necessitates the creation of initial datasets using rule-based classifiers, which may not capture the full variability and nuance present in real-world accident reports. This limitation can result in sub-optimal model performance, as deep learning approaches typically require extensive annotated corpora to generalize effectively. The vocabulary and structure of safety texts further complicate annotation efforts. Section headings and terminology can vary significantly across different reporting settings, making it difficult to establish consistent annotation guidelines and to ensure interoperability between datasets [5]. Domain-specific terms or keywords, such as those found in National Transportation Safety Board (NTSB) reports, are often not effectively identified by general extraction models, leading to gaps in the annotated data and reducing the utility of the resulting models for specialized applications [38]. This challenge is compounded by the fact that many existing text quantification methods, such as one-hot encoding, word2vec, and doc2vec, tend to ignore the contextual semantic relationships that are critical for understanding the intricacies of safety narratives. Another significant challenge arises from the need to construct both positive and negative samples for training robust question-answering systems. While positive samples, pairs of questions and their correct answers, are typically available, negative samples must be artificially generated by mismatching questions with unrelated answers. Ensuring a balanced dataset with sufficient negative samples is crucial for effective model training, yet this process is labor-intensive and requires careful curation to avoid introducing bias or noise. The annotation process is further complicated by the inherent ambiguity and fragmentation of safety texts, which often contain incomplete or context-dependent information that is difficult to interpret without domain expertise [11]. Manual annotation of safety texts is resource-intensive, demanding significant time and expertise from annotators who must possess both linguistic and domain-specific knowledge. The process is not only laborious but also prone to inconsistencies, as different annotators may interpret the same narrative differently, especially when the source of an accident or hazard is not explicitly stated [39]. Even advanced pre-trained language models, such as BERT, require fine-tuning on domain-specific data to achieve high accuracy, underscoring the importance of high-quality annotations for effective transfer learning [21][5]. Yugu et al. **0210406** indicate that domain-specific pretraining from scratch can outperform mixed-domain approaches, but this strategy is often infeasible in safety-critical domains due to the limited availability of annotated texts. Active learning frameworks offer a promising solution by iteratively selecting the most informative samples for annotation, thereby maximizing the efficiency of the annotation process and reducing the overall labeling burden [5]. However, the effectiveness of active learning is still constrained by the initial quality and representativeness of the seed data, as well as by the challenges associated with accurately quantifying uncertainty and selecting samples that will most benefit model performance [26]. The integration of structured and unstructured data, while enhancing the potential for extracting actionable insights, also increases the complexity of the annotation task, as annotators must navigate heterogeneous data formats and reconcile conflicting information [1]. Semantic understanding is another critical aspect, as many natural language processing tasks in the safety domain require models to grasp complex relationships between entities, such as the association between specific hazards and work activities [19][39]. The inability of general models to capture these nuanced relationships further highlights the necessity for domain-specific annotation strategies and specialized language models tailored to the safety context [38][19]. In summary, the annotation of safety texts is challenged by data scarcity, domain-specific vocabulary, the need for balanced positive and negative samples, the labor-intensive nature of manual labeling, and the limitations of general-purpose models in capturing the semantic richness of safety narratives. Addressing these challenges is essential for the development of effective, scalable question-answering systems capable of extracting critical safety insights from unstructured accident reports [5][11][38][39][26]**0210406**[21][19][1].

4 Data Sources and Corpus Construction

4.1 Types of Safety-Related Texts

4.1.1 Accident Investigation Reports

Accident investigation reports constitute a foundational data source for mining safety analysis, offering detailed, unstructured narratives that chronicle the sequence of events, risk factors, and contextual elements surrounding workplace incidents. These reports are typically generated following an accident and are designed to capture not only the immediate causes but also the underlying systemic issues that may have contributed to the event. The richness of information embedded within these documents makes them invaluable for extracting actionable safety insights, particularly when leveraging advanced natural language processing (NLP) techniques. The complexity and multifactorial nature of workplace accidents are often reflected in the narrative structure of these reports, which can include descriptions of environmental conditions, human factors, equipment failures, and organizational practices. Traditional approaches to analyzing such texts have relied heavily on manual coding and categorical analysis, which, while thorough, are inherently limited in scalability and subjectivity. The adoption of data-driven methodologies, such as those enabled by specialized language models, allows for a more granular and comprehensive examination of accident causality and contributing factors [1]. This methodological shift enhances the depth of accident investigations and supports the development of more effective, evidence-based safety interventions. Recent studies have demonstrated the utility of accident investigation reports as leading indicators for safety performance. By applying text classification and information extraction techniques, researchers have been able to identify critical information such as injury precursors, energy sources, accident types, and severity levels. This enables the prioritization of safety measures and the formulation of targeted accident prevention strategies. The ability to systematically analyze large volumes of such reports provides a strategic advantage in understanding the relationships between risk factors, accident typologies, and injury outcomes, which is essential for proactive safety management. The integration of advanced NLP models, such as SafetyBERT, further augments the analytical capabilities applied to accident investigation reports. These models are specifically tailored to the safety domain, allowing for more accurate extraction of domain-relevant entities and relationships compared to general-purpose language models. The performance of such specialized models has been shown to surpass that of baseline architectures, particularly in tasks involving the identification and classification of safety-critical information [2]. This underscores the importance of domain adaptation in NLP for high-stakes applications like mining safety. However, the effective utilization of accident investigation reports is not without challenges. The unstructured and heterogeneous nature of these texts, coupled with the often-limited availability of labeled data, poses significant obstacles to model training and evaluation. Manual annotation of large corpora is resource-intensive and time-consuming, necessitating the adoption of efficient data labeling strategies. Active learning frameworks have emerged as a promising solution, enabling the iterative selection of the most informative samples for annotation based on model uncertainty or confidence measures [5]. This approach minimizes the labeling burden while maximizing model performance, making it particularly well-suited for domains where expert-labeled data is scarce. The strategic use of accident investigation reports, in conjunction with advanced NLP and active learning, thus represents a significant advancement in the construction of robust, scalable safety analysis tools. By systematically extracting and synthesizing knowledge from these rich textual sources, it becomes possible to uncover latent patterns and risk factors that might otherwise remain obscured, ultimately contributing to the reduction of workplace accidents and the enhancement of industrial safety [1][2][5].

4.1.2 Near-Miss Narratives

Near-miss narratives constitute a critical subset of safety-related texts within mining operations, capturing incidents where hazardous events nearly resulted in injury, damage, or loss but were ultimately averted. These narratives are distinct from formal accident reports in that they often contain nuanced, context-rich descriptions of hazardous conditions, unsafe acts, or system failures that did not escalate to full-blown accidents. The value of near-miss narratives lies in their potential to reveal latent risks and precursors to more severe incidents, thereby enabling proactive hazard identification and mitigation strategies [2][18]. The unstructured nature of near-miss narratives presents unique challenges for automated extraction and analysis. Unlike structured incident logs, these texts are typically authored

by workers or supervisors in free-form language, often including colloquialisms, domain-specific jargon, and incomplete sentences. This variability complicates the direct application of general-purpose natural language processing (NLP) models, which may lack the specialized vocabulary and contextual understanding required to accurately interpret mining-specific terminology and operational scenarios **0210406**. For instance, general models may misinterpret technical terms or fail to recognize subtle indicators of risk embedded in the narrative, leading to reduced extraction accuracy and missed opportunities for safety improvement [3]. To address these challenges, domain-specific language models such as SafetyBERT are developed and fine-tuned on corpora rich in mining safety texts, including near-miss narratives. The authors of **0210406** indicate that pretraining language models from scratch on in-domain data enhances their ability to capture the semantics and structure of specialized texts, as the model's vocabulary and representations are tailored to the unique linguistic patterns of the mining domain. This approach mitigates the limitations observed when using models pretrained on general corpora, where rare or technical terms are often fragmented into subwords, diminishing the model's interpretability and downstream performance [3]. The annotation of near-miss narratives for supervised learning is a resource-intensive process, as it requires expert knowledge to accurately label the presence of hazards, contributing factors, and potential outcomes. Manual annotation at scale is often impractical, especially given the volume and diversity of narratives generated in large mining operations. To overcome this bottleneck, active learning strategies are employed, wherein the model iteratively selects the most informative or uncertain samples for annotation. This confidence-based sampling approach prioritizes near-miss narratives that are likely to yield the greatest improvement in model performance, thereby reducing the overall annotation burden while maintaining high-quality training data [9]. The gain in model performance achieved through active learning is particularly pronounced in smaller datasets, where each additional labeled instance can significantly enhance the model's ability to generalize [40]. Integrating near-miss narratives into the broader corpus for mining safety analysis not only enriches the dataset with diverse examples of hazardous scenarios but also supports advanced question-answering systems. These systems leverage the extracted insights from near-miss texts to provide timely, context-aware responses to safety-related queries, supporting decision-making and risk management at the operational level [11]. For example, during hazard identification (HAZID) activities, users can query the system for similar past near-miss events, enabling the retrieval of relevant narratives and the generation of actionable recommendations for mitigation [18]. The operability and comprehension of such systems are enhanced by the inclusion of near-miss data, as they reflect the real-world complexity and variability of mining environments. The combination of domain-adapted language models, active learning-driven annotation, and intelligent retrieval mechanisms forms a robust framework for harnessing the informational value of near-miss narratives. This approach not only addresses the technical challenges of unstructured text analysis but also aligns with the practical needs of high-stakes industrial settings, where timely and accurate safety insights are essential for preventing future incidents [11][40]**0210406**[9].

4.1.3 Safety Audits and Inspection Records

Safety audits and inspection records constitute a foundational component of safety-related textual data in industrial and mining environments. These documents are typically generated through systematic evaluations of operational sites, equipment, and procedures, aiming to identify hazards, assess compliance with safety standards, and recommend corrective actions. The textual content of such records is often unstructured, encompassing narrative descriptions, checklists, and evaluative comments, which presents unique challenges for automated information extraction and analysis [1]. Traditional approaches to analyzing safety audits and inspection records have relied heavily on manual review or basic keyword-based methods, such as Term Frequency–Inverse Document Frequency (TF-IDF), to identify recurring issues or trends [4]. However, these methods are limited in their ability to capture the nuanced context and latent patterns embedded within the narratives. For instance, frequency-based analyses may highlight the most commonly reported issues but can obscure less frequent yet critical sub-clusters, such as those related to specific operational failures or maintenance lapses. This limitation underscores the need for more sophisticated techniques capable of uncovering subtle correlations and integrating heterogeneous data sources, including both structured metadata and free-text narratives. Recent advancements in natural language processing (NLP) have enabled the development of more robust frameworks for processing and interpreting safety audits and inspection records. Embedding-based representations, enriched with metadata variables, allow for the construction of

comprehensive feature sets that better reflect the multifaceted nature of safety data. Dimensionality reduction techniques, such as Uniform Manifold Approximation and Projection (UMAP), facilitate the preservation of both local and global semantic structures within high-dimensional textual data, thereby enhancing the interpretability of extracted patterns. Subsequent clustering algorithms, like k-means, can then be applied to these reduced embeddings to identify latent groupings and recurring themes within the inspection records, offering deeper insights into accident causation and safety compliance [1][4]. The integration of advanced NLP models, such as SafetyBERT, further augments the analytical capabilities applied to safety audits and inspection records. These models leverage pre-trained language representations and can be fine-tuned on domain-specific corpora to improve their performance in extracting relevant entities, relations, and actionable insights from unstructured text [8][24]. The process typically involves pre-processing the raw records to standardize formats, remove noise, and segment narratives, followed by entity and relation extraction to populate structured knowledge graphs or databases. Such structured representations enable more effective querying, trend analysis, and decision support for safety management teams [38][4]. A significant challenge in constructing high-quality datasets from safety audits and inspection records is the resource-intensive nature of manual annotation, which is often required to generate labeled data for supervised learning tasks. Active learning strategies, which iteratively select the most informative samples for annotation based on model uncertainty or diversity, have been proposed to mitigate this bottleneck and accelerate the development of robust domain-specific models. By focusing annotation efforts on the most ambiguous or novel records, active learning can substantially reduce the amount of labeled data needed to achieve high model performance, making the analysis of large-scale safety audit corpora more feasible [41][17]. Moreover, the heterogeneity of safety audits and inspection records, spanning narrative descriptions, structured checklists, and quantitative measurements, necessitates the use of models and pipelines capable of integrating diverse data types. Classical models often struggle with this integration, leading to incomplete or fragmented analyses. In contrast, modern NLP pipelines can combine textual and structured data, enabling more holistic and scalable safety analysis tools suitable for high-stakes industrial environments [1]. The authors of Fan et al. [16] indicate that while machine learning approaches have been widely adopted for accident analysis, they often oversimplify the rich information contained in inspection records by reducing them to fixed numerical features, thereby losing event-level detail. Advanced NLP techniques, particularly those based on transformer architectures, offer a pathway to retain and exploit the full informational content of these records for more accurate and actionable safety insights [24][8]. In summary, safety audits and inspection records represent a complex yet invaluable source of data for mining safety analysis. The transition from traditional manual and frequency-based methods to advanced NLP-driven frameworks, incorporating domain-specific models, active learning, and integrated data representations, is essential for extracting critical safety insights and supporting proactive risk management in industrial settings [1][4][38].

4.1.4 Academic and Regulatory Documents

Academic and regulatory documents constitute a foundational component of safety-related text corpora, offering both theoretical and practical perspectives essential for comprehensive safety analysis. These documents are typically sourced from peer-reviewed journals, conference proceedings, regulatory agencies, and industry organizations, each contributing unique insights and data structures. The systematic collection of such heterogeneous data requires a multi-source extraction strategy, as outlined by Danish et al., who emphasize the necessity of targeting diverse sources to ensure both breadth and depth in domain coverage. This approach involves tailoring extraction methodologies to the specific accessibility, format, and structure of each source, thereby maintaining high data quality and relevance. Academic literature provides a rigorous, research-driven view of safety, often presenting novel methodologies, case studies, and meta-analyses that inform best practices and future research directions. Regulatory documents, on the other hand, encapsulate the legal and procedural frameworks governing safety standards, compliance requirements, and incident reporting protocols. The integration of these two types of documents enables the construction of a corpus that not only reflects the current state of scientific knowledge but also aligns with real-world operational and legal constraints [2]. The extraction and processing of academic and regulatory texts present several challenges. These documents are frequently unstructured or semi-structured, with varying terminologies, document layouts, and annotation conventions. Addressing these challenges necessitates the development of specialized extraction pipelines capable of handling diverse data formats, such as PDFs, HTML, and XML, as well as the

normalization of domain-specific vocabularies. The authors of indicate that moving beyond manual coding and categorical analysis is crucial for scalability and for capturing the multifactorial nature of workplace accidents. Automated extraction and text mining techniques, therefore, play a critical role in transforming raw documents into structured, analyzable data. Dimensionality reduction techniques are often employed to manage the high dimensionality inherent in textual data representations derived from academic and regulatory sources [1]. These methods facilitate efficient downstream processing, such as classification, clustering, and information retrieval, by reducing computational complexity while preserving essential semantic information. The necessity for such techniques becomes particularly pronounced when dealing with large-scale corpora, where the volume and diversity of documents can otherwise overwhelm traditional analytical approaches. Furthermore, the inclusion of regulatory documents ensures that the resulting corpus is not only academically robust but also practically applicable. Regulatory texts often contain standardized terminologies, incident classification schemes, and reporting templates that can be leveraged to harmonize data from disparate sources. This harmonization is vital for training domain-specific language models, such as SafetyBERT, which rely on consistent and representative input data to achieve high performance in specialized tasks. The comprehensive coverage achieved by integrating academic and regulatory documents supports the development of advanced NLP systems capable of extracting actionable safety insights from unstructured accident reports. By systematically curating and processing these texts, researchers can construct corpora that underpin scalable, data-driven safety analysis frameworks, ultimately enhancing the effectiveness of safety interventions in high-stakes industrial environments [2][1].

4.2 Data Collection and Preprocessing

4.2.1 Data Acquisition from Heterogeneous Sources

Data acquisition from heterogeneous sources is a foundational step in constructing a robust corpus for mining safety question-answering systems. The diversity of data sources is essential to capture the full spectrum of linguistic patterns, terminologies, and contextual nuances present in industrial safety narratives. This diversity, however, introduces significant challenges related to data collection, standardization, and integration, as each source may present unique formats, structures, and accessibility constraints [2]. For instance, incident reports, inspection logs, regulatory documents, and technical manuals often differ in their level of detail, language style, and metadata availability, necessitating tailored preprocessing strategies to harmonize the data. The process typically begins with the identification and extraction of relevant incident data from multiple repositories. In the context of aviation safety, for example, incident data from the Remotely Piloted Aircraft Systems (RPAS) category can be systematically extracted and processed to create a structured corpus. Each incident is represented as a row in a dataset, with columns corresponding to distinct fields such as event description, contributing factors, and outcomes. This structured representation facilitates subsequent annotation and analysis, enabling the creation of datasets in formats such as CSV, which are amenable to downstream machine learning workflows. Ensuring data quality and reliability is paramount, particularly when the data originates from disparate sources. Peer review and validation mechanisms, such as the assessment and release of individual incident reports, play a crucial role in verifying the accuracy and completeness of the collected data [18]. The integration of data from multiple sources also requires careful handling of inconsistencies, missing values, and conflicting information. Standardization protocols, including the normalization of terminology and the alignment of field definitions, are necessary to achieve a coherent and interoperable dataset. Domain adaptation further complicates data acquisition, as the language and context of safety-related texts can vary significantly across industrial sectors. The development of a comprehensive, domain-adaptive language model for occupational safety relies on the aggregation of textual data that reflects the operational realities of diverse environments, from construction sites to manufacturing plants [2]. This necessitates the inclusion of both structured and unstructured data, encompassing narrative accident reports, regulatory guidelines, and technical documentation. Traditional NLP techniques, such as Term Frequency–Inverse Document Frequency (TFIDF) and topic modeling, have been employed to extract salient features from unstructured safety narratives, while clustering algorithms like K-means facilitate the identification of thematic patterns within large text corpora [4]. The complexity of textual data representations in safety mining underscores the need for dimensionality reduction techniques. High-dimensional embeddings, while rich in information, can hinder efficient analysis and model training. Dimensionality reduction methods are thus employed to

distill the most informative features from the data, enhancing both the granularity and interpretability of subsequent analyses [1]. This approach moves beyond manual coding and categorical analysis, enabling scalable, data-driven investigations into the multifactorial nature of workplace accidents. Active learning strategies are particularly valuable in this context, as they enable the efficient annotation of large, heterogeneous datasets with minimal manual effort. By leveraging intelligent, confidence-based sampling, the annotation process can prioritize the most informative or uncertain examples, thereby maximizing the impact of limited labeling resources [42]. This is especially important given the resource-intensive nature of manual data annotation in specialized domains such as mining safety. The integration of heterogeneous data sources ultimately supports the development of more effective, evidence-based safety interventions. By systematically collecting, standardizing, and analyzing data from a wide array of sources, researchers can uncover complex causal relationships and emergent patterns that inform both model development and practical safety management strategies [1][2][4].

4.2.2 Text Normalization and Cleaning

Text normalization and cleaning are foundational steps in preparing unstructured accident reports for advanced natural language processing (NLP) tasks in mining safety. The inherent complexity and high dimensionality of textual data, especially in industrial safety contexts, necessitate robust preprocessing to ensure that downstream models such as SafetyBERT can extract meaningful and actionable insights [1]. Raw accident narratives often contain domain-specific jargon, inconsistent terminology, typographical errors, and a mixture of structured and unstructured information, all of which can impede effective model training and inference [2][1]. A critical aspect of normalization involves standardizing the vocabulary used within the corpus. In specialized domains like mining safety, generic language models may lack the necessary lexicon to accurately interpret technical terms and context-specific expressions. To address this, domain adaptation strategies have been employed, such as augmenting the original model vocabulary with terms extracted from the target corpus. For instance, Zhou et al. describe a method where a tokenizer is applied to the domain-specific corpus to identify unique words absent from the base vocabulary, thereby enriching the model's linguistic repertoire and improving its ability to process specialized content [29]. This approach is particularly relevant for mining safety, where the inclusion of technical terms related to equipment, hazards, and procedures is essential for precise information extraction. Tokenization itself is a crucial preprocessing step, as it segments raw text into discrete units (tokens) that can be effectively processed by transformer-based models like BERT. The process must account for the nuances of domain-specific language, ensuring that compound terms and abbreviations common in accident reports are preserved and correctly interpreted [34]. The use of advanced tokenizers, such as those provided by the HuggingFace Transformers library, facilitates this process and supports reproducibility in model development [43]. Beyond vocabulary and tokenization, text cleaning encompasses the removal of extraneous characters, normalization of case, correction of misspellings, and standardization of date and numerical formats. These operations reduce noise and variability in the data, which is particularly important when dealing with high-dimensional embeddings. As outlined in, dimensionality reduction techniques are often applied to text embeddings to manage complexity, but their effectiveness is contingent on the quality of the underlying text representation. Clean, normalized input ensures that embeddings capture the true semantic content of the reports rather than artifacts of inconsistent formatting or typographical errors. Integrating structured metadata with cleaned narrative text further enhances the analytical framework. By aligning categorical variables such as accident type, location, and time with normalized narrative embeddings, researchers can perform more comprehensive analyses that account for both contextual and categorical factors [1]. This integration supports both descriptive and predictive analytics, enabling the identification of major causes and trends in accident data. The challenges of manual data annotation in specialized domains underscore the importance of automated and semi-automated cleaning pipelines. Active learning frameworks, which iteratively select the most informative samples for annotation based on model confidence, benefit significantly from high-quality, normalized input. Effective preprocessing reduces the annotation burden by minimizing ambiguities and inconsistencies that could otherwise confound both human annotators and machine learning models [10]. In summary, text normalization and cleaning are indispensable for constructing a reliable and scalable corpus for mining safety question-answering systems. These processes ensure that domain-specific language is accurately represented, noise is minimized, and both structured and unstructured data are harmonized for advanced NLP analysis [2][1][29].

4.2.3 Section and Subsection Identification in Safety Documents

Section and subsection identification in safety documents is a foundational step in constructing a high-quality corpus for mining safety analysis. Safety documents, such as accident reports and regulatory guidelines, are typically composed of heterogeneous content, including narrative descriptions, structured tables, and embedded metadata. The extraction of meaningful sections and subsections from these documents is essential for downstream natural language processing (NLP) tasks, as it enables the alignment of textual evidence with specific safety concepts and facilitates targeted information retrieval [2][16]. The process begins with the acquisition of raw safety documents, which often arrive in unstructured or semi-structured formats. For instance, accident reports may contain free-text narratives interspersed with categorical fields, while regulatory documents might follow a hierarchical structure with explicit section headers. To transform these diverse sources into a unified corpus, preprocessing routines must first identify and segment the documents into coherent sections and subsections. This segmentation is not trivial, as inconsistencies in formatting, missing headers, and variable document templates can obscure the underlying structure [44][16]. Advanced NLP techniques, particularly those leveraging pretrained language models, have demonstrated efficacy in automating section and subsection identification. By training models on annotated corpora where section boundaries are explicitly marked, it becomes possible to recognize linguistic cues and formatting patterns indicative of section transitions. For example, models can be fine-tuned to detect header-like phrases, enumerate lists, or recognize shifts in narrative style that typically signal the start of a new section [45]. The use of domain-specific models, such as SafetyBERT, further enhances performance by incorporating knowledge of mining safety terminology and document conventions, which are often absent in general-purpose models [7]. In the context of mining safety, preprocessing pipelines may include steps such as text normalization, removal of extraneous formatting, and the application of regular expressions to extract candidate section headers. These candidate headers are then validated using machine learning classifiers or rule-based heuristics, which can be iteratively refined through active learning strategies. Active learning is particularly valuable in this setting, as it allows the model to query human annotators for ambiguous cases, thereby improving accuracy with minimal manual effort [5]. This approach is especially important given the resource-intensive nature of manual annotation in specialized domains. Once sections and subsections are identified, the resulting structure enables more granular analysis of safety incidents. For example, accident narratives can be linked to specific equipment types or injury categories, supporting fine-grained classification and risk assessment [2][16]. Furthermore, the structured segmentation of documents facilitates the creation of event-based datasets, where each event is associated with contextual metadata extracted from relevant sections. This organization is crucial for training and evaluating downstream models, such as those used for accident type classification or safety hazard question answering [11][16]. The integration of section and subsection identification into the preprocessing workflow also supports the development of scalable, automated safety analysis tools. By ensuring that each document is consistently segmented, it becomes feasible to batch process large volumes of safety reports, enabling efficient extraction of actionable insights for industrial safety management. The reliability and applicability of these tools are further enhanced by leveraging experience data and domain knowledge, which are systematically organized through the identified document structure [11]. In summary, the identification of sections and subsections in safety documents is a critical preprocessing step that underpins the construction of robust, domain-specific corpora for mining safety analysis. The combination of advanced NLP models, active learning, and domain expertise enables the efficient and accurate segmentation of complex documents, laying the groundwork for effective information extraction and decision support in high-stakes industrial environments [45][5][2][16].

4.2.4 Tokenization and Sequence Length Considerations

Tokenization and sequence length are fundamental considerations in the preprocessing pipeline for mining safety-related unstructured text, particularly when leveraging advanced language models such as SafetyBERT. Tokenization, the process of segmenting raw text into discrete units (tokens), directly influences how effectively a model can represent and process domain-specific language. In specialized domains like mining safety, the vocabulary often contains technical jargon, abbreviations, and compound terms that are not well-represented in general-purpose tokenizers. This mismatch can lead to suboptimal tokenization, where critical terms are fragmented into multiple subwords, potentially diluting semantic meaning and impairing downstream model performance [46][2]. The choice between using

a general-domain tokenizer and constructing a domain-specific vocabulary is non-trivial. El Boukkouri et al. [46] demonstrate that retraining a general model on specialized texts can yield performance comparable to training a model from scratch with a domain-specific vocabulary. This finding suggests that, while domain adaptation through continued pretraining is effective, careful attention must still be paid to the tokenization process to ensure that key safety-related terminology is preserved as intact tokens whenever possible. The authors of [2] indicate that the degree of specialization in the vocabulary impacts the model's ability to capture nuanced domain semantics, which is particularly relevant for mining safety narratives where precise terminology is critical. Sequence length, defined as the maximum number of tokens processed by the model in a single pass, imposes practical constraints on the representation of lengthy accident reports. Many industrial safety documents are verbose, containing detailed descriptions of incidents, environmental conditions, and corrective actions. If the sequence length is set too short, essential context may be truncated, leading to information loss and degraded model performance. Conversely, excessively long sequences increase computational demands and may introduce noise, especially if irrelevant sections are included [14][2]. Zhen and Yang [14] highlight that smaller, fine-tuned models can efficiently capture domain-specific details when the input is appropriately structured, underscoring the importance of balancing sequence length to maximize relevant content while minimizing unnecessary overhead. The interplay between tokenization and sequence length is further complicated by the subword-based tokenization schemes employed by models like BERT. Technical terms unique to mining safety may be split into multiple subwords, inflating the effective sequence length and reducing the amount of unique content that can be processed per input. This phenomenon necessitates a careful calibration of both the tokenizer and the maximum sequence length parameter during preprocessing. The work of Danish et al. [2] provides empirical evidence that model training efficiency is sensitive to sequence length, with shorter sequences enabling faster per-epoch training times but potentially at the cost of omitting critical context. Moreover, the presence of adversarial or artificially constructed sentences, as explored in [9], can further stress the tokenization and sequence length configuration. When such sentences are appended to paragraphs, the tokenizer must robustly handle both natural and synthetic language, and the sequence length must be sufficient to accommodate these additions without truncating original content. This scenario is particularly relevant in the context of evaluating model robustness and generalization. In practice, the preprocessing pipeline should include an analysis of the distribution of document lengths and the frequency of domain-specific terms to inform the selection of tokenizer and sequence length parameters. The integration of active learning strategies, as discussed in [5], can also assist in identifying edge cases where tokenization or sequence length limitations may hinder model performance, enabling targeted refinement of preprocessing steps. Smetana et al. [4] note that traditional NLP techniques, such as TFIDF and clustering, are sensitive to tokenization granularity, further emphasizing the need for domain-aware preprocessing in safety-critical applications. Ultimately, optimizing tokenization and sequence length is a balancing act that requires iterative experimentation and validation. The goal is to maximize the retention of critical safety information while ensuring computational efficiency and model robustness across diverse mining accident narratives [46][14][9][2].

4.3 Corpus Analysis

4.3.1 Linguistic Diversity and Technical Terminology

Linguistic diversity and the prevalence of technical terminology are defining characteristics of corpora constructed from mining safety accident reports. These documents, authored by a range of personnel with varying expertise and backgrounds, exhibit substantial heterogeneity in language use, narrative style, and the specificity of technical expressions. The natural language found in such reports is shaped by the operational context, regulatory requirements, and the need to convey nuanced details about abnormal events, leading to a corpus that is both rich in domain-specific vocabulary and variable in linguistic structure [47]. Mining safety narratives often integrate specialized terms that are not commonly encountered in general language corpora. This technical vocabulary encompasses equipment names, operational procedures, hazard classifications, and regulatory codes, which are essential for precise communication within the field. The integration of such terminology poses significant challenges for natural language processing models that are primarily trained on general-purpose datasets, as these models may lack the contextual understanding required to accurately interpret and extract relevant information from specialized texts [2][43]. The authors of [47] indicate that incident reports are crafted

to provide clarity and contextual detail, yet their accessibility is mediated by the reader's familiarity with the technical lexicon. The diversity in linguistic expression is further amplified by the inclusion of reports from different subdomains within mining, each with its own jargon and reporting conventions. For instance, accident narratives may vary in their use of abbreviations, acronyms, and colloquial expressions, reflecting the operational culture of specific mining environments. This heterogeneity necessitates careful corpus construction to ensure that the resulting dataset captures the full spectrum of language used in the field, thereby supporting the development of robust domain-specific language models [2][38]. To address the challenges posed by linguistic diversity and technical terminology, recent research has advocated for the incorporation of academic abstracts and regulatory documents alongside incident narratives in the training corpus. This approach exposes language models to both specialized and general linguistic constructions, mitigating the risk of catastrophic forgetting, whereby a model trained exclusively on domain-specific texts loses its ability to generalize to broader language tasks [2]. By maintaining a balance between technical specificity and general language patterns, the corpus supports the development of models that are both accurate in domain-specific tasks and resilient to shifts in linguistic context. The evolution of question-answering systems in safety-critical domains underscores the importance of capturing semantic nuances embedded in technical language. Advanced models such as SafetyBERT are designed to leverage the rich semantic information present in mining safety reports, but their effectiveness is contingent upon the quality and representativeness of the underlying corpus [15][41]. The use of transformer-based architectures, which excel at encoding contextual relationships, is particularly advantageous in this setting, as it enables the model to disambiguate technical terms based on surrounding context and to recognize patterns in the way safety incidents are described [19][41]. Furthermore, the process of manual data annotation in such linguistically diverse corpora is resource-intensive, as annotators must possess both linguistic proficiency and domain expertise to accurately label technical content. Active learning strategies have been proposed to alleviate this burden by prioritizing the annotation of samples that are most informative for model improvement, thereby maximizing the efficiency of the annotation process in the presence of complex technical terminology [15][9]. In summary, the construction of a mining safety corpus that faithfully represents linguistic diversity and technical terminology is foundational for the development of effective domain-specific question-answering systems. The interplay between specialized vocabulary, narrative variability, and the need for balanced linguistic exposure shapes both the challenges and opportunities in this area of research [47][2][38][43].

4.3.2 Statistical Overview of Document Types

A comprehensive statistical overview of document types within the constructed corpus is essential for understanding the representativeness and diversity of the data used in developing a domain-specific question-answering system for mining safety. The corpus comprises a heterogeneous mixture of structured and unstructured documents, with accident reports forming the core unstructured narrative component. These narratives are typically supplemented by structured metadata, such as accident type, location, time, and affected body part, which together enable a richer semantic representation of each incident. The integration of these two data modalities, narrative text and categorical metadata, facilitates more nuanced analyses and supports advanced natural language processing (NLP) techniques tailored to the safety domain. The unstructured accident narratives vary significantly in length, lexical diversity, and syntactic complexity. This variability is crucial for training robust language models, as it exposes the system to a wide range of linguistic patterns and domain-specific terminology. The structured metadata, on the other hand, provides categorical anchors that can be leveraged for supervised learning tasks and for enriching the semantic context of the narratives [1]. The distribution of document types within the corpus is therefore not uniform; rather, it reflects the operational realities of industrial safety reporting, where detailed textual descriptions are often accompanied by concise, structured descriptors. From a statistical perspective, the corpus can be characterized by several key metrics. The total number of documents, the proportion of narrative versus structured entries, and the frequency distribution of metadata categories (such as accident types or locations) are all critical parameters. For instance, in related domains, pre-training corpora for specialized language models have been quantified in terms of token counts, with some financial corpora reaching up to 3.3 billion tokens across various document types, including corporate reports and analyst transcripts [48]. While the mining safety corpus may not reach this scale, the principle of quantifying document types by token count, document length, and category frequency remains applicable. The presence of both

short, metadata-driven entries and longer, free-text narratives introduces challenges for model training, particularly in balancing the representation of rare versus common document types. This imbalance can affect the model's ability to generalize across different safety scenarios. The application of topic modeling techniques, such as Latent Dirichlet Allocation (LDA), has proven effective in uncovering latent thematic structures within unstructured text corpora, enabling the identification of recurring safety themes that may not be immediately apparent through manual inspection [1]. Such statistical analyses provide insights into the prevalence of specific incident types, the co-occurrence of certain risk factors, and the linguistic markers associated with high-severity events. The corpus construction process also involves rigorous data cleaning and preprocessing steps, which can significantly impact the final size and composition of the dataset. For example, subtle differences in data cleaning protocols have been shown to result in substantial variations in corpus size, as observed in the preparation of Wikipedia-based datasets for language model pre-training [34]. Ensuring consistency in preprocessing is therefore vital for maintaining the integrity of statistical analyses and for enabling reproducible research. In addition to accident reports, the corpus may include supplementary document types such as safety bulletins, regulatory guidelines, and technical manuals. Each of these contributes distinct linguistic and informational characteristics, further enriching the dataset. The inclusion of diverse document types supports the development of more versatile language models capable of handling a broad spectrum of safety-related queries. The authors of [49] indicate that large language models (LLMs) are benchmarked on a variety of tasks, including text classification and question answering, which necessitates exposure to multiple document genres during training. The statistical overview of document types thus serves as a foundational element in corpus analysis, informing both the design of NLP models and the interpretation of their outputs. By systematically quantifying the distribution and characteristics of each document type, researchers can identify potential biases, address data sparsity issues, and optimize active learning strategies for efficient annotation [41]. This approach ultimately enhances the scalability and effectiveness of safety analysis tools in high-stakes industrial environments.

4.3.3 Balance between Practical Narratives and Academic Content

The interplay between practical narratives and academic content is a central consideration in constructing a corpus for mining safety question-answering systems. Accident reports in mining are typically composed of unstructured textual narratives that encapsulate the sequence of events, contextual factors, and causal mechanisms underlying each incident. These narratives are rich in semantic content, often providing nuanced details that structured data fields may overlook, such as the specific environmental conditions, human factors, and equipment interactions that contributed to the event [1]. The unstructured nature of these narratives presents both an opportunity and a challenge: while they offer a wealth of information for extracting actionable safety insights, their variability and informality can complicate the application of standard natural language processing (NLP) techniques. Academic content, in contrast, is characterized by its structured presentation, formal language, and adherence to established terminologies. This structure facilitates the development and evaluation of NLP models, as it reduces ambiguity and enhances consistency across the corpus. However, academic texts may lack the immediacy and contextual richness found in practical narratives, potentially omitting critical incident-specific details that are essential for effective safety analysis [37]. The challenge, therefore, lies in harmonizing these two sources of information to construct a corpus that is both representative of real-world scenarios and amenable to rigorous computational analysis. The integration of practical narratives into the corpus necessitates advanced NLP techniques capable of handling the inherent variability and complexity of unstructured text. Embedding models such as Word2Vec, GloVe, and BERT have demonstrated the ability to capture contextual relations and syntactical meaning within texts, enabling the analysis of word similarity and semantic relationships even in less formal narrative data [4]. The use of pre-trained models like BERT, and its domain-specific adaptations such as SafetyBERT, further enhances the system's ability to extract relevant information from accident reports by leveraging contextual embeddings that are sensitive to the unique linguistic patterns present in mining safety narratives [37][7]. Domain-specific pre-training has been shown to significantly improve model performance on specialized tasks, as evidenced by the superior results of models like FinBERT and ArcheoBERTje in their respective domains. In the context of mining safety, pre-training on corpora that include both practical narratives and academic content allows the model to develop a nuanced understanding of the terminology, event sequences, and causal relationships that are unique to the

field [37]. This approach ensures that the model is not only capable of interpreting formal academic descriptions but also adept at extracting critical insights from the more variable and context-rich practical narratives. Active learning strategies play a crucial role in balancing the representation of practical and academic content within the corpus. By employing confidence-based sampling methods, the annotation process can be directed towards the most informative and uncertain examples, which often arise from the less structured practical narratives. This targeted approach to data annotation maximizes the efficiency of manual labeling efforts, ensuring that the resulting corpus captures the full spectrum of linguistic and contextual variability present in mining safety reports while minimizing the resource burden associated with large-scale manual annotation. The construction of a balanced corpus also involves careful consideration of class imbalance and the scarcity of labeled data, particularly for rare but critical incident types. The authors of [13] indicate that combining active learning with robust transformer architectures like BERT can mitigate the challenges posed by imbalanced datasets, enabling the model to generalize effectively across both common and rare event types. This is particularly important in high-stakes industrial environments, where the ability to accurately extract and interpret safety-critical information from both practical narratives and academic content can have a direct impact on risk mitigation and accident prevention. In summary, achieving an effective balance between practical narratives and academic content in the corpus is essential for the development of robust, domain-specific question-answering systems in mining safety. This balance ensures that the system is grounded in real-world incident data while maintaining the rigor and consistency required for reliable computational analysis. The integration of advanced NLP techniques, domain-specific pre-training, and active learning strategies collectively supports the creation of a corpus that is both comprehensive and scalable, enabling the extraction of actionable safety insights from the complex and heterogeneous data sources characteristic of the mining industry [4][37][1][13].

5 Natural Language Processing Techniques for Safety Texts

5.1 Transformer-Based Language Models

5.1.1 Architecture and Functionality

Transformer-based language models have become the foundation for advanced question-answering systems in specialized domains such as mining safety, where extracting actionable insights from unstructured accident reports is critical. The architecture of these models is characterized by their ability to process sequential data through self-attention mechanisms, enabling the capture of complex dependencies within textual narratives. At the core, the transformer architecture operates by converting raw textual data into numerical representations, a process known as embedding, which is essential for any downstream natural language processing (NLP) task [6]. This embedding step ensures that the semantic and syntactic properties of the text are preserved in a format suitable for computational modeling. The functionality of transformer-based models, such as SafetyBERT, is built upon the pre-training and fine-tuning paradigm. Initially, a general-purpose language model like BERT is either further pre-trained on domain-specific corpora or trained from scratch using a specialized vocabulary tailored to the mining safety context [50]. This adaptation allows the model to internalize the unique terminology and linguistic patterns present in accident reports, which are often absent in general corpora. The process involves multi-task self-supervised pre-training, where the model learns from various objectives simultaneously, followed by supervised fine-tuning on annotated datasets relevant to safety question answering. The authors of [51] indicate that this multi-stage training pipeline enhances the model's ability to generalize across different language understanding tasks within the financial domain, and by analogy, similar strategies are effective for mining safety texts. A significant challenge in deploying such models in specialized fields is the scarcity of labeled data, as manual annotation of safety incidents is both time-consuming and resource-intensive [24]. To address this, the integration of active learning strategies is crucial. In this approach, the model iteratively selects the most informative or uncertain samples from a pool of unlabeled accident reports, which are then prioritized for annotation. This confidence-based sampling not only accelerates the learning process but also ensures that the annotation effort is focused on examples that are likely to yield the greatest improvement in model performance. According to [14], specialized domain adaptation, when combined with intelligent data selection, leads to superior performance compared to relying solely on general-purpose models or random sampling. The architecture of the question-answering system is further enhanced

by incorporating a classification layer, typically a linear layer with a softmax activation, which maps the contextualized embeddings produced by the transformer to specific answer categories or spans within the text [50]**0210406**. The training objective often involves minimizing a cross-entropy loss function, which aligns the model's predictions with the ground truth labels. This setup allows the system to intelligently select the most relevant answers to safety-related queries, leveraging both the contextual understanding provided by the transformer and the discriminative power of the classification layer [11]**0210406**. In practical deployments, such as safety hazard management systems, the transformer-based question-answering framework is not limited to dialog-based interactions. It can also visualize common hazards, output historical records, and automate the creation of management lists, thereby streamlining safety analysis and reducing the burden on human experts [11]. The efficiency and precision of these models have been demonstrated in related domains, where models like MiniLM achieve high F1 and exact match scores on benchmark datasets, indicating their capability to accurately extract relevant information from complex texts [35]. The process of embedding, model pre-training, active learning, and intelligent answer selection collectively form a robust architecture for mining safety question-answering systems. By leveraging domain-specific language models and adaptive data sampling strategies, these systems can scale to large volumes of unstructured reports while maintaining high accuracy and interpretability, which is essential for high-stakes industrial environments [6][51][50][14].

5.1.2 BERT and Its Domain Adaptations

BERT (Bidirectional Encoder Representations from Transformers) has established itself as a foundational architecture in natural language processing, particularly due to its ability to capture bidirectional context and semantic relationships within text. Its transformer-based design enables the model to process input sequences in parallel, leveraging self-attention mechanisms to encode rich contextual information. However, while BERT achieves strong performance on a variety of general NLP tasks, its effectiveness can be limited when applied directly to highly specialized domains such as mining safety, where domain-specific terminology and context are prevalent [28][52]. To address these limitations, researchers have explored domain adaptation strategies for BERT. One prominent approach involves pre-training or fine-tuning BERT on corpora that are representative of the target domain. For instance, models like BioBERT and SciBERT have demonstrated that adapting BERT to biomedical and scientific texts, respectively, leads to significant improvements in downstream tasks within those fields [28]. The process typically involves either continued pre-training on domain-specific unlabeled data or fine-tuning on labeled datasets relevant to the application area. This adaptation allows the model to internalize domain-specific vocabulary, phraseology, and semantic nuances that are absent or underrepresented in general corpora. The effectiveness of domain adaptation is further highlighted by comparative studies. For example, El Boukkouri et al. [46] indicate that training a model directly on a specialized corpus with a tailored vocabulary can outperform approaches that merely re-train a general BERT model on domain data, as seen in the case of medical text analysis. This suggests that the selection of vocabulary and the nature of the pre-training data are critical factors in maximizing model performance for specialized tasks. In addition to vocabulary and corpus adaptation, the development of domain-specific BERT variants has proliferated across various fields. Models such as FinBERT for financial texts, PatentBERT for patent literature, and SentiLR for sentiment analysis in specific contexts exemplify this trend [28]. Liu et al. [51] demonstrate that simultaneous training on both general and domain-specific corpora, as implemented in FinBERT, enables the model to capture a broader spectrum of language knowledge and semantic information, enhancing robustness and effectiveness in financial applications. The computational demands of training and deploying large transformer models like BERT have also motivated research into model compression and efficiency improvements. Techniques such as model pruning, quantization, parameter sharing, and knowledge distillation have led to the creation of lighter variants like DistilBERT, TinyBERT, and ALBERT, which retain much of the original model's performance while reducing resource requirements [28]. These advancements are particularly relevant for industrial applications where computational resources may be constrained. Despite these advances, challenges remain in determining the optimal adaptation strategy. Pearce et al. note that the performance gains from self-training models such as BERT and XLNet are influenced by factors like training set size and hyperparameter tuning, and that BERT itself was found to be undertrained in its original formulation. RoBERTa, an improved training methodology for BERT, addresses some of these issues by optimizing training procedures and leveraging larger datasets, resulting

in superior performance on a range of tasks [52]. Bosley et al. [17] further show that RoBERTa, when fine-tuned on specific classification tasks, can outperform even advanced generative models like GPT-3 in domain-specific applications. The adaptation of BERT for sentence-pair tasks, as described by Mohamed Hassan et al. [39], involves concatenating label descriptions with narrative text and feeding the sequence into the model. This approach enables the extraction of aggregate representations that are well-suited for classification and inference tasks in specialized domains. In summary, the evolution of BERT and its domain adaptations underscores the importance of aligning model architecture, vocabulary, and training data with the specific requirements of the target application. The proliferation of domain-specific BERT variants, coupled with advances in model efficiency and training methodologies, has significantly expanded the applicability of transformer-based models to specialized fields such as mining safety, biomedical research, and finance [46][28][51][52][39][17].

5.1.3 Comparison with Lightweight and Large-Scale Models

The comparison between lightweight and large-scale transformer-based language models is central to the design of effective question-answering systems for specialized domains such as mining safety. Large-scale models, exemplified by BERT and its derivatives, have demonstrated remarkable capabilities in extracting semantic and contextual information from unstructured texts due to their deep architectures and extensive pretraining on massive corpora [35][29]. For instance, the BERT Base model consists of 12 encoder layers and is trained on tasks such as masked language modeling and next sentence prediction, enabling it to capture nuanced language patterns [35]. However, these models are computationally intensive, requiring significant resources for both training and inference, which can limit their practical deployment in real-time or resource-constrained environments [53]. Lightweight models, on the other hand, are designed to reduce computational overhead while maintaining competitive performance. Approaches such as the Reranker, which fine-tunes transformer models using localized contrastive estimation loss, exemplify efficient strategies for information retrieval and question answering without the need for deep, resource-heavy architectures [19]. These models often employ parameter sharing, reduced layer counts, or distilled architectures to achieve faster inference and lower memory consumption. The trade-off, however, is that lightweight models may not capture the full complexity of domain-specific language, especially in highly specialized fields where subtle contextual cues are critical for accurate interpretation [24]. Hybrid architectures that combine the strengths of both paradigms have emerged as promising solutions. For example, integrating BERT with BiGRU and self-attention mechanisms has been shown to significantly improve text feature extraction, outperforming models that rely solely on CNN, LSTM, or BERT subnets. This fusion leverages the representational power of large-scale models while introducing architectural efficiencies that can mitigate some of the computational burdens. The authors of [11] indicate that such hybrid models achieve notable gains in accuracy, highlighting the potential for tailored architectures in specialized applications. Tokenization strategies also play a crucial role in the effectiveness of both lightweight and large-scale models. BERT and similar models utilize subword tokenization to manage vocabulary size and handle out-of-vocabulary terms, which is particularly important when dealing with technical jargon or rare terminology in safety reports [46]. However, the reliance on subword units can introduce challenges in accurately representing domain-specific concepts, suggesting that further customization of tokenization schemes may be necessary for optimal performance in specialized domains [29]. Fine-tuning strategies are another axis of comparison. Large-scale models benefit from extensive pretraining but often require substantial labeled data for effective domain adaptation **0210406**. In contrast, lightweight models can be more amenable to rapid adaptation with limited data, especially when combined with active learning techniques that prioritize the annotation of high-uncertainty samples [20]. This approach not only reduces the annotation burden but also accelerates the convergence of the model to domain-specific tasks, making it a practical choice for environments where labeled data is scarce and annotation is costly [24]. The scalability of these models is also influenced by their parameter optimization strategies. For instance, the PAL-BERT model demonstrates that careful selection and tuning of internal parameters, such as the number of layers, can yield substantial improvements in question-answering performance without incurring the full computational cost of the largest available models [53]. This underscores the importance of model architecture search and hyperparameter optimization in balancing performance and efficiency. In multilingual and cross-domain contexts, lightweight models may struggle to generalize across languages or domains without additional adaptation, whereas large-scale models pretrained on diverse corpora can offer broader coverage but at the expense of increased resource requirements

[8][17]. The choice between these approaches must therefore consider the specific demands of the application, including the complexity of the language, the availability of computational resources, and the criticality of accurate information extraction in safety-critical settings. Ultimately, the integration of specialized language models such as SafetyBERT with intelligent sampling and active learning frameworks represents a pragmatic synthesis of these approaches. By leveraging the deep contextual understanding of large-scale models and the efficiency of lightweight architectures, it is possible to construct scalable, high-performance question-answering systems tailored to the unique challenges of mining safety analysis [20][11][19][35].

5.2 Domain Adaptation Strategies

5.2.1 Domain-Specific Pre-Training

Domain-specific pre-training is a critical step in adapting general-purpose language models to specialized domains such as mining safety, where the linguistic characteristics and terminology differ substantially from those found in open-domain corpora. The process involves further training of a pre-existing language model, such as BERT or RoBERTa, on a large corpus of domain-relevant texts, thereby enabling the model to internalize the unique vocabulary, phraseology, and contextual nuances present in safety-related documents [2][51][17]. This adaptation is essential because general models, while powerful, often lack the capacity to accurately interpret and extract meaningful information from highly specialized texts without additional exposure to domain-specific data [24]. The methodology for domain-specific pre-training typically begins with the selection or construction of a representative corpus. In the context of mining safety, this may include unstructured accident reports, safety bulletins, and regulatory documents. The model is then subjected to continual pre-training using objectives such as masked language modeling (MLM), which encourages the model to predict masked tokens based on their context, thereby reinforcing its understanding of domain-specific semantics [2][51]. Danish et al. [2] outline that this continual training phase is instrumental in transforming a general-purpose model into one that is attuned to the occupational safety domain, enhancing its ability to capture subtle linguistic cues that are critical for accurate information extraction. The effectiveness of domain-specific pre-training has been demonstrated across various fields. For instance, Liu et al. [51] show that simultaneous training on both general and domain-specific corpora enables models like FinBERT to capture both broad linguistic knowledge and specialized semantic information. This dual exposure is particularly beneficial in domains where the language exhibits both general and highly technical characteristics. Similarly, Bosley et al. [17] report that additional pre-training of RoBERTa on domain-specific datasets leads to improved performance on downstream tasks compared to models trained solely on general corpora. However, the benefits of domain-specific pre-training are not always uniform across all evaluation metrics. Yuan [24] observes that further pre-training BERT on a large financial domain corpus yields only marginal improvements in certain metrics such as Precision@1 and may even result in slight decreases in others like NDCG. This suggests that while domain adaptation generally enhances model relevance, the magnitude of improvement can depend on the alignment between the pre-training corpus and the target task. Another important consideration is the preservation of general language capabilities during domain adaptation. Overfitting to the domain corpus can lead to a loss of generalization, making the model less effective on texts that blend domain-specific and general language. To mitigate this, approaches such as mixed precision training and the use of both general and domain-specific corpora during pre-training have been employed, as demonstrated by Liu et al. [51]. This strategy ensures that the model retains a balance between specialized knowledge and general linguistic competence. In the context of mining safety, domain-specific pre-training enables models like SafetyBERT to more effectively extract critical safety insights from unstructured accident reports. By internalizing the patterns and terminology unique to safety narratives, the model can identify causal factors, categorize accident types, and support proactive safety interventions [4]. The integration of domain-adapted models with active learning frameworks further enhances their utility, allowing for efficient model improvement with minimal labeled data by focusing annotation efforts on the most informative samples. The process of domain-specific pre-training is thus a cornerstone of effective domain adaptation strategies in NLP for safety-critical applications. It bridges the gap between general language understanding and the specialized requirements of industrial safety analysis, enabling the development of robust, interpretable, and scalable tools for extracting actionable knowledge from complex, unstructured texts [2][51][17][24].

5.2.2 Fine-Tuning for Specialized Tasks

Fine-tuning large language models for specialized tasks is a critical step in adapting general-purpose NLP architectures to the nuanced requirements of mining safety question-answering systems. The process involves taking a pre-trained model, such as BERT or its derivatives, and further training it on domain-specific data to enhance its ability to extract relevant information from unstructured safety reports. This approach addresses the inherent limitations of generic models, which often lack the contextual sensitivity required for high-stakes industrial applications [2]. The effectiveness of fine-tuning is closely tied to the quality and specificity of the training data. When models are exposed to domain-specific corpora, they develop a more refined understanding of the terminology, phraseology, and contextual cues unique to mining safety documentation. Kanakarajan et al. [3] demonstrate that pretraining and fine-tuning on domain-specific text, coupled with a tailored vocabulary, significantly improves the model's ability to capture contextual representations, leading to more accurate and reliable extraction of safety insights. A central challenge in this process is the scarcity of annotated data in specialized domains. Manual annotation is resource-intensive and often impractical at scale. To mitigate this, active learning strategies are employed, where the model iteratively selects the most informative samples for annotation based on confidence scores or uncertainty estimates. This targeted approach ensures that the limited annotation budget is utilized efficiently, accelerating the improvement of model performance with minimal labeled data [2][19]. The integration of active learning with fine-tuning creates a feedback loop that incrementally enhances the model's domain adaptation capabilities. The technical implementation of fine-tuning typically involves concatenating the question and candidate answer sequences, applying tokenization, and leveraging segment embeddings to distinguish between the two inputs. The model is then trained to predict the relevance of each candidate answer, outputting a probability score that reflects its confidence in the answer's correctness [24]. This architecture allows the model to learn subtle relationships between questions and context passages, which is essential for accurate answer selection in complex safety scenarios. Comparative studies have shown that fine-tuned BERT-family models, such as RoBERTa and SafetyBERT, consistently outperform zero-shot and in-context learning approaches, particularly in classification and question-answering tasks within specialized domains [17][2]. Bosley et al. [17] highlight that the performance gains from fine-tuning are substantial, underscoring the continued relevance of this strategy even as more advanced generative models become available. Furthermore, the use of domain-adapted models like SafetyBERT, which are specifically optimized for safety-related texts, yields superior results compared to more generic architectures [2]. Parameter optimization during fine-tuning is another crucial aspect. For instance, the PAL-BERT model introduces strategies such as gradient accumulation and reduced-precision training to enhance efficiency and performance during the adaptation process [53]. These optimizations enable the handling of larger datasets and more complex model architectures without prohibitive computational costs. In specialized domains like mining safety, the combination of fine-tuning with intelligent sampling and parameter optimization forms a robust framework for extracting actionable insights from unstructured text. This approach not only improves the accuracy of question-answering systems but also ensures scalability and adaptability to evolving safety documentation practices [2][3][53][19]. The iterative refinement enabled by active learning and domain-specific fine-tuning is essential for developing reliable, high-performance NLP tools in safety-critical environments.

5.2.3 Handling Technical Jargon and Contextual Semantics

Handling technical jargon and contextual semantics in mining safety texts presents unique challenges for natural language processing systems. The specialized vocabulary found in accident reports and safety documentation often diverges significantly from general language, necessitating tailored approaches for effective information extraction. General-domain language models, such as the original BERT, utilize vocabularies and tokenization strategies optimized for broad coverage, which can result in suboptimal representation of domain-specific terms when applied to specialized corpora. This mismatch leads to fragmented tokenization of technical terms, reducing the model's ability to capture their semantic integrity and, consequently, diminishing downstream task performance. To address these limitations, domain adaptation strategies have been developed that involve constructing custom vocabularies and retraining or fine-tuning language models on domain-specific corpora. For instance, the creation of a medical WordPiece vocabulary and its application to medical texts has demonstrated

improved tokenization and representation of specialized terminology compared to general-domain vocabularies [46]. Similarly, in the context of clinical natural language processing, the introduction of customized vocabularies has been shown to further enhance the performance of BERT-based models, as these vocabularies better align with the linguistic characteristics of the target domain [29]. The process of adapting the vocabulary is not merely a technical adjustment but a critical step in ensuring that the model can accurately encode and interpret the nuanced meanings embedded in technical jargon. Beyond vocabulary adaptation, the semantic complexity of mining safety texts requires models to capture contextual dependencies that extend across sentences and document structures. Transformer-based architectures, such as those underlying SafetyBERT, are particularly well-suited for this task due to their self-attention mechanisms, which enable the modeling of long-range dependencies and intricate semantic relationships [19]. Fine-tuning these models on domain-specific datasets allows them to internalize the contextual semantics unique to safety narratives, including the interplay between environmental factors, human behaviors, and equipment attributes that contribute to accident causation [14]. The authors of [50] indicate that domain-adapted models, such as FinBERT in the financial sector, outperform their general-purpose counterparts on semantic tasks, underscoring the importance of both vocabulary and contextual adaptation. The challenge of acquiring sufficient labeled data for effective domain adaptation is compounded by the resource-intensive nature of manual annotation in specialized fields. Active learning strategies offer a solution by prioritizing the annotation of data points where the model exhibits low confidence, thereby maximizing the informational gain from each labeled instance [42]. This approach is particularly advantageous in high-stakes environments like mining safety, where expert annotation is costly and time-consuming. By iteratively refining the model with the most informative examples, active learning accelerates the acquisition of domain-specific knowledge and enhances the model's ability to handle technical jargon and complex semantics [7]. Visualization and interpretability techniques, such as t-SNE and UMAP, play a complementary role by enabling researchers to inspect the structure of high-dimensional embeddings generated by adapted models [1]. These methods facilitate the identification of clusters corresponding to technical terms and semantic themes, providing insights into how well the model captures domain-specific concepts. Such interpretability is crucial for validating the effectiveness of adaptation strategies and for uncovering latent relationships within safety texts that may inform risk mitigation efforts. In summary, the effective handling of technical jargon and contextual semantics in mining safety texts hinges on a combination of vocabulary customization, contextual fine-tuning, active learning for efficient annotation, and interpretability tools for model validation. These strategies collectively enable the development of robust, domain-adapted NLP systems capable of extracting actionable safety insights from complex, unstructured data [46][1][19][14][50][7][29][42].

5.3 Question Answering System Design

5.3.1 Extractive QA Frameworks

Extractive question answering (QA) frameworks have become essential in mining safety applications, where the extraction of precise, contextually relevant information from unstructured accident reports is critical for effective safety analysis. These frameworks are designed to identify and extract specific spans of text from documents that directly answer a posed question, rather than generating new text or summaries. The effectiveness of extractive QA in safety-critical domains is closely tied to the ability of the underlying language models to comprehend domain-specific terminology, event sequences, and nuanced contextual cues present in accident narratives [14][2]. Traditional extractive QA systems often rely on deep learning architectures such as bidirectional long short-term memory (BiLSTM) networks, sometimes enhanced with attention or coattention mechanisms. These models encode both the question and the context passage, learning to align relevant segments of the text with the information need expressed in the question. For instance, a stacked BiLSTM with coattention can capture intricate relationships between questions and answers, leveraging both cosine similarity and Euclidean distance metrics to assess the relevance of candidate answer spans [35]. Such architectures have demonstrated strong performance in general QA tasks, but their effectiveness in specialized domains like mining safety is limited by their dependence on large, annotated datasets and their lack of domain adaptation [24]. The emergence of pre-trained language models (PLMs) such as BERT and its domain-adapted variants has significantly advanced extractive QA capabilities. PLMs are trained on vast corpora and can be further fine-tuned on domain-specific data to enhance their understanding of specialized

language and context. In mining safety, models like SafetyBERT are tailored to capture the unique linguistic patterns and event structures found in accident reports, enabling more accurate extraction of critical safety insights [2]. The top layers of BERT, in particular, have been shown to be highly effective for text classification and QA tasks, especially when further pre-training is performed on in-domain data, which mitigates issues like catastrophic forgetting and boosts performance even with limited labeled examples [40]. A key challenge in developing extractive QA frameworks for mining safety is the scarcity of annotated data, as manual labeling of accident reports is both time-consuming and resource-intensive [16][29]. To address this, active learning strategies have been integrated into the QA pipeline. Active learning leverages model uncertainty to intelligently select the most informative samples for annotation, thereby maximizing the efficiency of the labeling process. For example, confidence-based sampling strategies prioritize instances where the model is least certain, ensuring that each annotated example contributes maximally to model improvement [15]. This approach is particularly valuable in high-stakes industrial environments, where rapid adaptation to new data and evolving safety concerns is essential. The integration of large language models (LLMs) into extractive QA frameworks further enhances their utility. LLMs can process unstructured crash narratives, preserving the sequential flow of events and capturing complex multi-vehicle interactions that are often lost in tabular representations. Moreover, LLMs are capable of generating natural language explanations for their predictions, making the extracted answers more interpretable for safety practitioners who may lack technical expertise [14][36]. This interpretability is crucial for the adoption of automated QA systems in operational safety analysis. Recent advancements also include machine-guided report generation, where models like ChatGPT are employed to transform unstructured accident narratives into structured templates, facilitating efficient data extraction and reducing the need for extensive human annotation [16]. Such approaches ensure that the richness and completeness of crash contexts are retained, supporting more comprehensive safety analyses. In summary, extractive QA frameworks for mining safety leverage a combination of advanced NLP architectures, domain-adapted PLMs, and active learning strategies to efficiently extract actionable insights from unstructured accident reports. These systems address the dual challenges of domain specificity and limited labeled data, providing scalable and interpretable solutions for safety-critical applications [2][14][16][40][35].

5.3.2 Generative QA Approaches

Generative question answering (QA) approaches have gained significant traction in recent years, particularly for extracting actionable knowledge from unstructured safety texts such as mining accident reports. Unlike extractive QA, which identifies answer spans directly from the input text, generative QA models synthesize responses, often leveraging large-scale pre-trained language models that can generate coherent and contextually relevant answers based on the input query and supporting documents [27]. This generative capability is especially valuable in safety-critical domains, where nuanced interpretation and synthesis of information are required to support decision-making. The core of generative QA systems is typically a transformer-based architecture, such as BERT or its derivatives, which encodes both the question and the context into dense vector representations. These representations are then used to generate answers, either by selecting relevant spans or by producing free-form text that addresses the query [54]. The use of BERT and similar models has demonstrated strong performance in various QA benchmarks, with models trained on large-scale datasets like SQuAD showing robust generalization to new tasks [27]. However, when applied to specialized domains such as mining safety, these general-purpose models often struggle to capture domain-specific terminology and context, leading to suboptimal performance [24]. To address these limitations, domain-adapted models such as SafetyBERT have been developed. These models are pre-trained or fine-tuned on corpora specific to the target domain, enabling them to better understand the unique linguistic patterns and knowledge structures present in safety texts [7][2]. For instance, Kumar et al. [7] demonstrate that domain-specific BERT variants outperform general scientific language models on specialized evaluation sets, highlighting the importance of tailored pre-training for effective generative QA in technical fields. A significant challenge in deploying generative QA systems for mining safety is the scarcity of labeled data. Annotating accident reports with question-answer pairs requires expert knowledge and is resource-intensive. Transfer learning and active learning strategies have been proposed to mitigate this bottleneck. Transfer learning leverages knowledge from large, general datasets and adapts it to the target domain, reducing the need for extensive manual annotation [24]. Active learning further enhances efficiency by intelligently selecting the most informative samples for annotation, often based on model

uncertainty or confidence scores, thereby maximizing the impact of limited labeling resources [19]. The architecture of generative QA systems often incorporates mechanisms to align question and context representations, such as coattention or self-attention layers, which facilitate the modeling of complex relationships between queries and supporting texts [35][54]. These mechanisms enable the model to focus on relevant portions of the input, improving answer quality and relevance. Additionally, recent advances have explored the integration of Siamese networks and sentence embedding techniques to capture semantic similarity between questions and candidate answers, further enhancing the system's ability to generate accurate and context-aware responses [19][35]. Evaluation of generative QA models in the safety domain typically involves both quantitative metrics, such as exact match and F1 scores, and qualitative assessments of answer relevance and utility for safety professionals [27][33]. The ability to generate precise, contextually appropriate answers is crucial for supporting hazard identification, risk assessment, and the development of preventive measures in high-stakes industrial environments [2]. Moreover, the deployment of such systems as virtual assistants or learning tools can facilitate knowledge transfer and support training initiatives, as they are capable of interpreting complex queries and providing tailored explanations based on extensive safety documentation [22]. In summary, generative QA approaches represent a promising direction for mining safety analysis, combining advanced natural language processing techniques with domain adaptation and efficient data annotation strategies. The integration of specialized language models, intelligent sampling methods, and sophisticated attention mechanisms enables the development of scalable, effective tools for extracting critical safety insights from unstructured texts [24][27][19][7].

5.3.3 Integration of Domain Knowledge in QA

Integrating domain knowledge into question answering (QA) systems for mining safety is essential for achieving accurate and contextually relevant responses, especially when dealing with highly specialized and technical accident reports. General-purpose language models, while powerful, often struggle to interpret the nuanced terminology and unique linguistic patterns present in safety-critical domains. This limitation arises because pre-trained language models are typically exposed to broad, general corpora during training, which do not capture the specific vocabulary, phraseology, and implicit knowledge embedded in mining safety texts. As a result, direct application of such models to specialized domains can lead to suboptimal performance, particularly in extracting actionable safety insights. To address these challenges, domain adaptation strategies are employed, where models are either fine-tuned on domain-specific corpora or explicitly infused with structured domain knowledge. For instance, specialized models like SafetyBERT are developed by further pre-training or fine-tuning on mining safety datasets, enabling the model to internalize the semantics and context of safety-related language [51][52]. This process enhances the model's ability to disambiguate technical terms, recognize incident patterns, and understand regulatory or procedural references that are prevalent in mining safety documentation. The integration of domain knowledge can also be achieved through the incorporation of structured information during pre-training. Techniques such as masking contiguous spans of domain-relevant terms or introducing span boundary objectives, as seen in models like SpanBERT and StructBERT, allow the model to learn relationships between key concepts and their contextual boundaries [28]. By predicting masked spans based on their boundaries, the model is encouraged to capture the structural dependencies and semantic roles that are critical in safety narratives. This approach is particularly beneficial in mining safety, where the relationships between entities (e.g., equipment, hazards, actions) are central to understanding the causality and consequences of incidents. Another promising direction involves the integration of external knowledge bases or ontologies specific to mining safety. By embedding structured safety knowledge into pre-trained language models (PLMs), the system can achieve deeper contextual reasoning and improved interpretability. This integration supports the model in handling rare or emerging terminology, as well as in reasoning about complex scenarios that may not be well represented in the training data. The authors of [2] suggest that future research should further explore the cross-domain transferability and zero-shot or few-shot adaptation capabilities of such models, especially in low-resource settings where annotated data is scarce. Active learning frameworks further enhance the integration of domain knowledge by iteratively selecting the most informative or uncertain samples for annotation and model refinement [25]. In the context of mining safety QA, confidence-based sampling strategies can prioritize accident reports or narrative segments where the model exhibits low certainty, thereby focusing human annotation efforts on the most challenging and impactful cases. This targeted approach not only reduces the annotation burden but also accelerates

the model's acquisition of domain-specific knowledge, leading to more robust and scalable QA systems [15]. The design of input representations also plays a crucial role in leveraging domain knowledge. For example, concatenating context and question pairs and carefully tokenizing them ensures that the model receives coherent and contextually rich input sequences, which is vital for accurate answer extraction in technical domains [52]. Setting appropriate sequence lengths and handling special tokens further optimize the model's ability to process long and information-dense safety documents. Moreover, the use of text-generation methods, where the model generates answers token by token conditioned on the question, context, and previously generated tokens, allows for flexible and context-aware response generation [20]. This sequential approach is particularly suited for mining safety QA, where answers often require synthesizing information from multiple parts of a report and reasoning over complex event chains. Pre-processing steps, such as standardizing terminology and ensuring that datasets remain representative of real-world safety incidents, are also critical for maintaining the integrity of domain knowledge throughout the QA pipeline [18]. Careful attention to context during pre-processing prevents the loss of meaning and supports the model in generalizing to new or unseen scenarios. Finally, the integration of domain knowledge is not limited to model architecture and training strategies but extends to the design of evaluation benchmarks and annotation guidelines. Constructing domain-specific benchmarks and providing clear annotation protocols ensure that the QA system is evaluated on tasks that reflect real-world safety analysis needs, thereby aligning model development with practical requirements [12][17]. In summary, the effective integration of domain knowledge in mining safety QA systems is achieved through a combination of domain-adaptive pre-training, structured knowledge incorporation, active learning, tailored input representations, and rigorous data pre-processing. These strategies collectively enable the development of robust, interpretable, and scalable QA tools capable of extracting critical safety insights from unstructured accident reports [20][51][25][2][15][52][17][28][18].

6 Efficient Model Development for Mining Safety QA

6.1 Active Learning Frameworks

6.1.1 Principles of Confidence-Based Query Selection

Confidence-based query selection is a foundational principle in active learning frameworks, particularly when optimizing the annotation process for domain-specific question-answering systems in high-stakes environments such as mining safety. The core idea is to prioritize the selection of data samples for annotation based on the model's uncertainty or confidence in its predictions. By focusing annotation efforts on instances where the model exhibits low confidence, the framework can maximize the informational gain from each labeled example, thereby accelerating model improvement while minimizing manual annotation costs [20]. In the context of mining safety QA, unstructured accident reports often contain complex, domain-specific language and nuanced event descriptions. General-purpose language models may struggle to interpret these intricacies, leading to unreliable predictions and highlighting the necessity for targeted data selection strategies. Confidence-based query selection addresses this by systematically identifying those samples where the model's output probability distribution is either flat or ambiguous, indicating uncertainty. These samples are then prioritized for expert annotation, ensuring that the most informative and challenging cases are incorporated into the training set [28][20]. The implementation of confidence-based selection typically involves quantifying the model's uncertainty using metrics such as entropy, margin sampling, or the difference between the top two predicted probabilities. For example, in a classification setting, the entropy H of the predicted probability vector $\mathbf{p}$ for a sample can be computed as

$$H(\mathbf{p}) = -\sum_{i=1}^{C} p_i \log p_i$$

where C is the number of classes and p_i is the predicted probability for class i. Samples with higher entropy are considered less confidently classified and are thus prime candidates for annotation [28]. This approach is particularly effective in domains where the cost of annotation is high and the available labeled data is limited, as is often the case in mining safety. Recent advances in large language models (LLMs) and their application to QA systems have further refined confidence-based query selection. For instance, LLMs can be prompted to self-assess their reasoning and flag answers where logical consistency or factual accuracy is questionable, effectively generating an internal confidence score [20].

This self-evaluation mechanism can be integrated into the active learning loop, allowing the system to autonomously identify and request expert input on ambiguous or complex cases. Such strategies have been shown to improve both the efficiency and the robustness of domain-specific QA systems, especially when combined with specialized models like SafetyBERT that are tailored to the linguistic characteristics of mining safety reports [28][20]. Moreover, the integration of confidence-based sampling with domain-adapted models leverages the strengths of both approaches. Domain-specific pretraining, as demonstrated in biomedical and financial QA systems, enhances the model's baseline performance on specialized texts, while active learning ensures that annotation resources are allocated where they are most needed **0210406**[24]. This synergy is crucial for developing scalable and effective QA tools in environments where safety-critical decisions depend on accurate information extraction from unstructured data. The iterative nature of confidence-based active learning also facilitates continuous model refinement. As new accident reports are generated and processed, the system can dynamically update its confidence estimates and select new queries for annotation, maintaining high performance even as the underlying data distribution evolves. This adaptability is essential in industrial settings where operational conditions and reporting practices may change over time. In summary, confidence-based query selection is a scientifically grounded strategy that enhances the efficiency of active learning frameworks for mining safety QA. By systematically targeting low-confidence predictions for expert review, the approach accelerates model improvement, reduces annotation costs, and supports the development of reliable, domain-specific language models capable of extracting actionable safety insights from complex, unstructured reports [20]**0210406**[28][24].

6.1.2 Seed Set Selection and Iterative Annotation

Seed set selection and iterative annotation are fundamental components in the design of an active learning framework for mining safety question-answering systems. The initial seed set, typically a small subset of unlabeled accident reports, is crucial as it forms the foundation upon which the model begins its learning process. In pool-based active learning, a pool of unlabeled data is available, and instances are drawn from this pool for annotation, as outlined by Zhang et al.. The selection of the seed set must be performed with care to ensure that it is representative of the broader data distribution, thereby enabling the model to capture the diversity of accident scenarios encountered in mining safety contexts. The iterative annotation process is characterized by a loop in which the current model is used to select new instances from the unlabeled pool that are most informative for annotation. This selection is often guided by a confidence-based sampling strategy, where the model identifies instances about which it is least certain. By focusing annotation efforts on these uncertain cases, the framework maximizes the information gain from each annotation cycle, leading to more rapid improvements in model performance with fewer labeled examples. This approach is particularly advantageous in domains such as mining safety, where manual annotation is resource-intensive and expert annotators are scarce. Cost-sensitive querying strategies can further refine the selection process by considering both the reliability of annotators and the informativeness of instances. In practice, this may involve estimating the trustworthiness of different annotators and dynamically assigning instances to those with higher reliability, thereby improving the quality of the labeled data. Additionally, requiring multiple annotations per instance and consolidating these annotations can help mitigate individual annotator biases and enhance label accuracy [26]. The iterative nature of the annotation process allows for continuous refinement of the model. After each annotation cycle, the model is retrained with the newly labeled data, and its parameters are updated accordingly. This loop continues until the model achieves satisfactory performance or the annotation budget is exhausted. The use of neural embedding models, such as SafetyBERT, facilitates the extraction of dense vector representations from unstructured accident reports, enabling more effective clustering and similarity-based selection of instances for annotation. Embedding-based clustering can reveal underlying patterns and causal factors in accident data, which can inform the selection of diverse and representative seed sets as well as guide subsequent annotation rounds [1]. Active learning frameworks that incorporate iterative annotation and intelligent seed set selection are particularly well-suited for high-stakes industrial environments. They enable the efficient development of specialized language models by strategically allocating annotation resources to the most impactful data points. This is especially relevant in mining safety, where the consequences of model errors can be severe, and the need for robust, domain-adapted models is paramount. The integration of multitask learning approaches, as discussed by Danish et al. [2], can further enhance the framework by capturing interconnected effects

across different safety-related tasks, allowing for more nuanced and comprehensive safety insights. The iterative annotation process is also influenced by the need to avoid overfitting and ensure generalization. Khattab et al. highlight the importance of deterministic data splits to prevent the model from seeing the same question during both training and retrieval, which is critical for maintaining the integrity of the evaluation process. By carefully managing the flow of data through the annotation and training pipeline, the framework can achieve a balance between model accuracy and annotation efficiency [33]. In summary, the combination of representative seed set selection, confidence-based iterative annotation, cost-sensitive querying, and embedding-based clustering forms a robust foundation for active learning in mining safety QA systems. This approach not only addresses the challenges of limited labeled data and domain specificity but also supports the scalable and effective extraction of actionable safety insights from complex, unstructured accident reports [1][26][33][2].

6.1.3 Balancing High-Confidence and Low-Confidence Samples

Balancing high-confidence and low-confidence samples is a central consideration in the design of active learning frameworks for mining safety question-answering systems. The effectiveness of active learning hinges on the strategic selection of data points for annotation, with the aim of maximizing model improvement while minimizing annotation costs. In the context of mining safety, where unstructured accident reports contain critical information, the challenge is to identify which samples will most efficiently enhance the specialized language model, such as SafetyBERT, for extracting actionable safety insights. Active learning typically leverages model uncertainty as a proxy for informativeness, prioritizing samples where the model exhibits low confidence in its predictions. This approach is grounded in the observation that uncertain samples are likely to reside near the decision boundary, and their annotation can lead to significant updates in the model's parameters. Grießhaber et al. outline that Bayesian approximations of model uncertainty can be effectively used to select such low-confidence samples for manual labeling, thereby improving the efficiency of the annotation process [42]. Zhu et al. further demonstrate that pool-based active learning algorithms, including least confidence, margin sampling, and entropy-based methods, are particularly effective in multitask scenarios, as they systematically target samples that the model finds ambiguous or challenging [15]. However, an exclusive focus on low-confidence samples can introduce biases and limit the diversity of the training set. If the active learning process continually selects only the most uncertain samples, the model may overfit to rare or atypical cases, neglecting the broader distribution of the data. This is especially problematic in high-stakes domains like mining safety, where both common and rare accident scenarios must be accurately understood. Rose et al. indicate that meaningful subdivisions exist even within well-defined clusters of accident data, each characterized by distinct metadata values. This underscores the necessity of maintaining a representative sample pool that captures the full spectrum of accident scenarios, not just those that are ambiguous to the model. Balancing high-confidence and low-confidence samples thus becomes a nuanced task. While low-confidence samples drive learning by challenging the model, high-confidence samples serve as anchors, reinforcing the model's understanding of well-established patterns. Incorporating a mix of both types ensures that the model does not drift away from the core distribution of the data while still being exposed to novel or difficult cases. This balance is particularly relevant when deploying specialized models like SafetyBERT, which must generalize across diverse accident narratives and structured metadata [1]. The computational strategies for achieving this balance often involve hybrid sampling techniques. For instance, entropy-based methods can be combined with random sampling to ensure that both uncertain and representative samples are included in the annotation queue. Additionally, the use of joint models that capture relationships between different features or tasks can provide marginal distributions that inform more sophisticated sampling strategies, as highlighted by Zhu et al. [15]. These approaches enable the active learning framework to remain both efficient and robust, adapting to the evolving needs of the mining safety QA system. Another consideration is the resource-intensive nature of manual annotation in specialized domains. By judiciously selecting a balanced set of high- and low-confidence samples, the annotation effort can be directed where it is most impactful, reducing redundancy and accelerating model convergence [42][15]. This is particularly important in mining safety, where expert annotators are often required, and the cost of annotation is high. The integration of narrative and structured data further complicates the sampling process. Rose et al. emphasize that combining narrative embeddings with metadata enhances the granularity of accident analysis and supports the development of predictive models. Therefore, the sampling strategy must account for both the narrative complexity and the

structured attributes of each report, ensuring that the selected samples are informative across multiple dimensions. In summary, the process of balancing high-confidence and low-confidence samples in active learning frameworks for mining safety QA is a dynamic optimization problem. It requires careful consideration of model uncertainty, data diversity, annotation costs, and the interplay between narrative and structured information. By leveraging Bayesian uncertainty estimates, pool-based algorithms, and hybrid sampling strategies, the framework can efficiently guide the annotation process, resulting in a scalable and effective safety analysis tool tailored to the unique challenges of the mining industry [1][42][15].

6.2 Annotation and Human-in-the-Loop Processes

6.2.1 Guidelines for Expert Annotation

Expert annotation is a cornerstone for developing robust question-answering systems in the mining safety domain, especially when leveraging advanced NLP models such as SafetyBERT. The annotation process must be meticulously designed to ensure that the resulting labeled data is both accurate and representative of the complex, domain-specific phenomena present in unstructured accident reports. Given the high-stakes nature of safety-critical environments, annotation guidelines should prioritize clarity, consistency, and domain relevance to maximize the utility of expert input for downstream model training and evaluation. A primary consideration is the selection and training of annotators. Annotators should possess substantial domain expertise in mining safety, as well as familiarity with the linguistic and contextual nuances of accident narratives. This expertise is essential for accurately interpreting the multifaceted factors, such as temporal, environmental, vehicular, and human elements, that contribute to incident outcomes [14]. The annotation guidelines must therefore provide explicit definitions and examples for each target entity, relation, or answer span, reducing ambiguity and promoting uniformity across annotators [28]. To further enhance annotation quality, regular calibration sessions and inter-annotator agreement assessments are recommended, allowing for the identification and resolution of discrepancies in interpretation [12]. The annotation protocol should address the inherent challenges of unstructured and variable narrative data. Accident reports often exhibit significant linguistic variability across temporal and jurisdictional dimensions, which can hinder consistent labeling. To mitigate this, guidelines should instruct annotators to focus on the semantic integrity of the text, emphasizing the preservation of factual content while standardizing terminology and phrasing where appropriate. This approach not only improves the coherence of annotated data but also facilitates more effective downstream processing by NLP models [14][1]. Given the resource-intensive nature of manual annotation, integrating active learning strategies can substantially improve efficiency. By employing a confidence-based sampling approach, the annotation process can prioritize instances where the model exhibits the greatest uncertainty, thereby maximizing the informational value of each labeled example [36]. This targeted annotation strategy is particularly advantageous in domains where expert time is limited and annotation costs are high. Annotators should be provided with clear instructions on how to handle ambiguous or borderline cases, including the option to flag instances for further review or to indicate when no suitable answer is present, as is common in advanced QA systems [27]. The annotation workflow should also incorporate mechanisms for iterative feedback and continuous improvement. After initial annotation rounds, model predictions can be reviewed by experts to identify systematic errors or biases, informing subsequent refinements to both the annotation guidelines and the model itself. This human-in-the-loop paradigm not only enhances model performance but also reduces the burden of exhaustive manual review by leveraging the complementary strengths of automated and expert-driven analysis [12][4]. To ensure that annotated data supports the development of scalable and generalizable safety analysis tools, guidelines should encourage annotators to capture the full spectrum of relevant factors and interactions described in accident narratives. This includes not only explicit mentions of causes and outcomes but also implicit contextual cues that may influence incident dynamics [14]. The annotation schema should be sufficiently flexible to accommodate the complexity of real-world safety events, while remaining structured enough to support reliable extraction and modeling by NLP systems [1][14]. Finally, the annotation process should be documented in detail, including the rationale for schema design, instructions provided to annotators, and procedures for quality control. Transparent documentation facilitates reproducibility and enables future researchers to build upon the annotated datasets with confidence [40]. By adhering to these guidelines, expert annotation can serve as a foundation for the development of effective, domain-adapted question-answering systems

that advance safety management in mining and related industries [1][14][12].

6.2.2 Annotation Tools and Workflow Optimization

Annotation tools and workflow optimization are essential components in the development of efficient, high-quality mining safety question-answering systems. The process of annotating unstructured accident reports is inherently resource-intensive, particularly in specialized domains where expert knowledge is required to accurately label complex safety events and causal factors. Traditional annotation workflows, which often involve manual labeling from scratch, can be prohibitively slow and costly, especially when large-scale datasets are needed to train advanced language models such as SafetyBERT [29][8]. To address these challenges, integrating intelligent annotation tools that leverage model-assisted workflows has become increasingly important. One effective strategy is the use of pre-annotation, where the model generates initial predictions or highlights relevant spans within the text, which are then reviewed and corrected by human annotators. This approach reduces the cognitive load on annotators, as they can focus on validating or refining model suggestions rather than performing exhaustive manual labeling. Zhang et al. [26] state that pre-annotation enables annotators to interact more closely with the model, allowing them to select or adjust the model's top predictions, thereby accelerating the creation of high-quality gold-standard annotations. Active learning further enhances annotation efficiency by prioritizing the selection of data samples that are most informative for model improvement. In this paradigm, the model identifies instances where its confidence is lowest or where disagreement among model committee members is highest, and these samples are then presented to annotators for labeling. Zhu et al. [15] outline that strategies such as least confidence sampling and query-by-committee (QBC) can be employed to systematically select the most valuable samples from the unlabeled data pool, ensuring that annotation efforts are concentrated where they yield the greatest impact on model performance. Workflow optimization also involves the integration of annotation tools with downstream evaluation and feedback mechanisms. For example, annotation platforms can be designed to tag each report or paragraph with unique identifiers, facilitating traceability and enabling efficient evaluation of model outputs against annotated ground truth [38]. This traceability is particularly important in high-stakes environments like mining safety, where the interpretability and reliability of extracted insights are critical. Another aspect of workflow optimization is the adaptation of annotation tools to domain-specific requirements. General-purpose annotation platforms may not adequately support the specialized terminology and event structures encountered in mining safety reports. As such, customization of annotation interfaces, including the incorporation of domain-specific taxonomies and entity types, is necessary to ensure accurate and consistent labeling [29][8]. The use of domain-adapted language models, such as those pre-trained or further pre-trained on mining or safety-related corpora, can also improve the quality of pre-annotations and reduce the burden on human annotators [24][8]. The combination of model-assisted annotation, active learning, and domain-specific tool customization creates a synergistic workflow that maximizes annotation efficiency and quality. By iteratively refining the model with targeted human feedback, the annotation process becomes more scalable and sustainable, enabling the rapid development of robust mining safety QA systems. This approach not only accelerates dataset creation but also ensures that the resulting models are well-calibrated to the nuances of mining safety language and reporting practices [26][15][38][29]. In summary, optimizing annotation tools and workflows through model-assisted pre-annotation, active learning-driven sample selection, and domain-specific customization is fundamental to the efficient development of high-performance mining safety question-answering systems. These strategies collectively reduce annotation costs, improve data quality, and facilitate the deployment of interpretable and reliable safety analysis tools in industrial settings [26][15][29][38].

6.2.3 Reducing Annotation Burden with Active Learning

Reducing the annotation burden is a central challenge in developing domain-specific question-answering systems for mining safety, especially when leveraging advanced NLP models such as SafetyBERT. Manual annotation of unstructured accident reports is not only time-consuming but also requires significant domain expertise, making it a costly bottleneck in the creation of high-quality datasets for supervised learning. Active learning (AL) offers a promising solution by strategically selecting the most informative samples for annotation, thereby maximizing model improvement per labeled instance and minimizing overall annotation effort. The core principle of active learning is to iteratively identify

data points about which the model is least confident, and then prioritize these for human annotation. This approach is particularly effective in specialized domains like mining safety, where the distribution of critical incidents is often skewed and rare events carry disproportionate importance. By focusing annotation efforts on ambiguous or uncertain cases, AL ensures that the model rapidly acquires knowledge about the most challenging aspects of the data, leading to faster convergence and improved generalization. In practical terms, a typical AL workflow begins with a small, seed set of labeled accident reports. An initial model is trained on this dataset, and then used to predict labels for a larger pool of unlabeled reports. The model's confidence in its predictions is quantified, often using metrics such as the least confidence score or entropy. Instances with the lowest confidence are selected for annotation in each iteration. This pool-based sampling scenario, as described in [5], allows for efficient allocation of annotation resources, as only the most informative samples are labeled by human experts before retraining the model. The integration of AL with advanced language models such as BERT or its domain-specific variants (e.g., SafetyBERT) further enhances the efficiency of this process. Pre-trained models can encode complex contextual relationships within the text, enabling more accurate identification of ambiguous cases that would benefit most from expert annotation [7][11]. The ensemble learning strategies described in demonstrate that combining rule-based NLP algorithms with AL-driven sampling can bootstrap the initial dataset, providing a robust foundation for subsequent model refinement. Moreover, the iterative nature of AL aligns well with human-in-the-loop processes, where domain experts are engaged only when their input is most valuable. This not only reduces the overall annotation workload but also ensures that expert time is spent on the most impactful cases. The workflow outlined in exemplifies this approach, where embedding vectors are constructed from the initially labeled data, and subsequent model training is guided by the active selection of uncertain instances. The effectiveness of AL in reducing annotation burden is further supported by the observation that, in specialized domains, the marginal utility of each additional labeled instance is higher when selected via confidence-based strategies compared to random sampling. This is particularly relevant in mining safety, where the cost of misclassification can be substantial, and the diversity of incident types necessitates targeted learning [5][11]. Additionally, the use of pre-trained language models in conjunction with AL mitigates the need for extensive manual feature engineering or the development of bespoke rule sets for each new dataset. As outlined by PankajKumar et al. [7], simultaneous learning of tokenization, named entity recognition, and text parsing during pretraining allows the model to capture semantic connections between properties directly from unlabeled text, further reducing the reliance on annotated data. In summary, active learning represents a scalable and effective strategy for minimizing annotation requirements in the development of mining safety QA systems. By leveraging model uncertainty to guide annotation, integrating advanced language models for contextual understanding, and employing iterative human-in-the-loop workflows, it is possible to construct high-performance, domain-specific QA tools with significantly reduced manual labeling effort [5][7][11].

6.3 Model Evaluation and Continuous Improvement

6.3.1 Performance Metrics for QA Systems

Performance metrics are fundamental for assessing the effectiveness and reliability of question-answering (QA) systems, especially in high-stakes domains such as mining safety. The selection and interpretation of these metrics directly influence the iterative development and deployment of such systems. In QA, the most widely adopted metrics include Exact Match (EM), F1-score, and, in some cases, semantic similarity measures such as cosine similarity and BLEU score, each capturing different aspects of system performance. The Exact Match metric quantifies the proportion of predictions that exactly correspond to the ground truth answers. This metric is particularly stringent, as it requires the predicted answer to match the reference answer word-for-word, without allowance for paraphrasing or minor variations [21]. While EM provides a clear-cut measure of correctness, it may underestimate the system's utility in cases where semantically equivalent but lexically different answers are produced. To address the limitations of EM, the F1-score is commonly employed. The F1-score is the harmonic mean of precision and recall, calculated at the token level between the predicted and reference answers. Precision measures the fraction of predicted tokens that are correct, while recall assesses the fraction of ground truth tokens that are retrieved by the model. The F1-score thus balances the trade-off between these two aspects, offering a more nuanced evaluation, especially in scenarios where partial correctness

is valuable **0210406**[21]. For instance, in named entity recognition (NER) tasks within QA pipelines, entity-level F1-scores are used to evaluate the extraction of critical information from unstructured text **0210406**. Semantic similarity metrics, such as cosine similarity and BLEU score, provide additional perspectives on answer quality. Cosine similarity evaluates the closeness of vector representations of predicted and reference answers, capturing semantic overlap even when lexical forms differ [21]. BLEU score, originally developed for machine translation, measures the n-gram overlap between predicted and reference answers, rewarding both precision and fluency. These metrics are particularly useful when the QA system is expected to generate natural language responses rather than extract spans verbatim from the input text. In specialized domains like mining safety, the presence of technical jargon and domain-specific expressions poses unique challenges for standard metrics. For example, the use of static masking and domain-adapted language models, such as KTL-BERT, has demonstrated improvements in F1-score, indicating that tailored pretraining strategies can enhance the extraction of relevant entities and relations from technical documents. The authors of [8] indicate that such domain-specific adaptations can yield F1-score improvements ranging from approximately 1% to over 15% compared to generic models, highlighting the importance of metric-driven evaluation in guiding model selection and refinement. Beyond answer accuracy, QA systems in safety-critical applications must also consider answerability. The ability to correctly identify unanswerable questions is crucial to prevent the dissemination of misleading or incomplete information. One approach involves augmenting the model output with a binary indicator reflecting whether a question is answerable, thereby allowing the system to abstain from providing answers when appropriate [43]. This strategy is evaluated using metrics such as answerability accuracy and the rate of false positives and negatives, which are essential for maintaining trust in automated safety analysis tools. Efficiency metrics, such as inference speed and computational resource usage, are also relevant in the context of scalable QA systems. For instance, models like DistilBERT achieve a favorable balance between performance and efficiency, retaining approximately 95% of BERT's accuracy while reducing model size and computational requirements by 40% and 60%, respectively [17]. Such trade-offs are particularly pertinent when deploying QA systems in resource-constrained industrial environments. Continuous improvement of QA systems relies on the systematic application of these metrics throughout the model development lifecycle. Active learning frameworks, which iteratively select the most informative samples for annotation based on model uncertainty, leverage performance metrics to prioritize data labeling and monitor gains in accuracy and robustness [40][13]. For example, BERT-SL has been shown to outperform classical BERT and active learning baselines by up to 4.72% and 13.30%, respectively, as measured by standard evaluation metrics. This demonstrates the value of integrating metric-driven feedback into the active learning loop to maximize annotation efficiency and model performance. In summary, the rigorous application of performance metrics such as Exact Match, F1-score, semantic similarity, answerability accuracy, and efficiency measures is indispensable for the development, evaluation, and continuous improvement of QA systems tailored to mining safety. These metrics not only quantify system effectiveness but also inform strategic decisions regarding model architecture, domain adaptation, and data annotation strategies, ultimately supporting the deployment of reliable and scalable safety analysis solutions [40]**0210406**[43][8][17][21].

6.3.2 Iterative Training and Feedback Loops

Iterative training and feedback loops are fundamental to the continuous improvement of domain-specific question-answering (QA) systems in mining safety. The process begins with an initial model, such as SafetyBERT, which is fine-tuned on a limited set of annotated accident reports. However, the scarcity and high cost of manual annotation in specialized domains like mining safety necessitate strategies that maximize the utility of each labeled instance **0210406**. Active learning addresses this by employing intelligent sampling strategies that select the most informative or uncertain examples for annotation, thereby optimizing the annotation effort and accelerating model improvement [13]. In each iteration, the model is exposed to new data points, often selected based on confidence scores or uncertainty estimates. These data points are then annotated and incorporated into the training set, allowing the model to learn from its mistakes and adapt to previously unseen patterns [40]**0210406**. This feedback loop is not only efficient but also crucial for handling the evolving nature of safety narratives, where new types of incidents or terminology may emerge over time. The iterative process ensures that the model remains up-to-date and robust against distributional shifts in the data [20]. The effectiveness of this approach is further enhanced by leveraging domain-specific language mod-

els. For instance, SafetyBERT, which is tailored to the mining safety context, can better capture the nuances of accident narratives compared to general-purpose models. The embedding construction in such models encodes both positional and structural information, which is critical for understanding complex safety reports. As the model iteratively incorporates feedback from newly labeled data, its internal representations become increasingly aligned with the domain-specific semantics required for accurate QA [2][45]. Model evaluation is tightly integrated into the feedback loop. After each training cycle, the model's performance is assessed using metrics such as exact match scores or F1 scores on a held-out validation set. This evaluation guides the selection of subsequent data points for annotation, focusing on areas where the model exhibits the greatest uncertainty or lowest performance. The authors of indicate that hyperparameter optimization during fine-tuning can further enhance the gains achieved through iterative feedback, as each cycle provides new insights into optimal training configurations. Visualization tools, such as attention pattern visualizations in BERT-based architectures, offer additional feedback by revealing how the model interprets linguistic relationships within the text [7]. These insights can inform both model developers and annotators, highlighting potential sources of error or bias in the data and guiding targeted improvements in subsequent iterations. Iterative feedback loops also facilitate the adaptation of QA systems to complex, multi-faceted queries that are common in mining safety analysis. Traditional machine learning models often struggle with such tasks, but the combination of active learning and advanced language models enables the system to incrementally build competence in handling nuanced, multi-step reasoning required for extracting actionable safety insights [12]. The integration of information retrieval components within the feedback loop further enhances the system's ability to extract relevant knowledge from large, unstructured corpora. As outlined in [20], the correctness of the final answer serves as a reward signal, guiding the selection and refinement of query rewrites and retrieval strategies. This dynamic adjustment ensures that the QA system remains effective even as the underlying data and user requirements evolve. In summary, iterative training and feedback loops, underpinned by active learning and domain-specific modeling, are essential for developing scalable, high-performance QA systems in mining safety. These mechanisms enable efficient use of limited labeled data, continuous adaptation to new information, and sustained improvements in model accuracy and reliability **0210406**[13][2][7].

6.3.3 Error Analysis and Model Uncertainty

Error analysis and model uncertainty are central to the robust evaluation and iterative enhancement of domain-specific question-answering (QA) systems in mining safety. In high-stakes environments, such as mining, the consequences of misclassification or missed critical information can be severe, making it essential to systematically identify, quantify, and address sources of error and uncertainty in model predictions. A key challenge in mining safety QA arises from the inherent complexity and variability of unstructured accident reports. These documents often contain domain-specific terminology, ambiguous phrasing, and rare event descriptions, which can confound general-purpose language models. Safety-domain models like SafetyBERT have demonstrated superior performance in classifying critical and rare categories, particularly in data-constrained and class-imbalanced settings, where general models frequently fail to detect such cases [2]. This observation underscores the necessity of domain adaptation and specialized pretraining to reduce systematic errors associated with out-of-domain or underrepresented classes. Fine-tuning optimization plays a significant role in minimizing model errors. The process of hyperparameter selection, as shown by PankajKumar et al. [7], can substantially impact the exact match scores and overall QA performance. Iterative experimentation with over a hundred configurations for each BERT-based model revealed that even minor adjustments in learning rate, batch size, or sequence length can lead to notable improvements or degradations in model accuracy. This sensitivity highlights the importance of rigorous validation and error tracking during model development. Active learning frameworks further contribute to error reduction by strategically selecting the most informative samples for annotation. By employing confidence-based sampling strategies, the system prioritizes instances where the model exhibits high uncertainty, thereby focusing human labeling efforts on ambiguous or challenging cases [15]. This targeted approach not only accelerates performance gains but also exposes systematic weaknesses in the model, such as confusion between similar safety hazard categories or misinterpretation of context-dependent cues. The integration of batch processing for large-scale safety hazard analysis, as described by Dan Tiana et al. [11], enables efficient identification of problematic cases across massive datasets. Users can rapidly export and review answers to specific questions, facilitating the manual inspection of model outputs and the

detection of recurring error patterns. Such batch evaluation is particularly valuable for uncovering edge cases and rare failure modes that may not be apparent in small-scale testing. Model uncertainty estimation is crucial for risk-aware deployment in industrial settings. Safety-domain models, including SafetyBERT and safetyALBERT, have shown resilience in maintaining performance advantages under class imbalance and limited data scenarios [2]. However, quantifying uncertainty remains a nontrivial task. Confidence scores derived from softmax outputs or alternative uncertainty metrics can be used to flag low-confidence predictions for human review, thereby reducing the likelihood of critical errors propagating into operational decisions. The iterative nature of model evaluation and improvement is further supported by the use of relevance-guided supervision and fine-grained interaction mechanisms, as demonstrated in ColBERT-QA [33]. These techniques enable the model to refine its retrieval and extraction capabilities based on weak heuristics or relevance feedback, systematically reducing retrieval errors and enhancing answer precision. Error analysis also benefits from the complementary strengths of hybrid QA architectures. The combination of knowledge graph-based QA (KGQA) and deep learning-based QA (DLQA) modules, as shown by Agarwal et al., allows the system to leverage structured domain knowledge alongside contextual language understanding. This synergy helps mitigate errors arising from either module in isolation, leading to improved accuracy and recall in mining safety applications [38]. Domain-specific pretraining, as evidenced by the PureMechBERT models, enables the capture of unique contextual relations within technical texts that general-purpose models may overlook [7]. This embedded domain knowledge can be harnessed to reduce semantic errors and improve the interpretability of model predictions, particularly for nuanced safety-related queries. Continuous error analysis, supported by intelligent sampling, batch evaluation, and uncertainty quantification, forms the backbone of an effective and scalable mining safety QA system. By systematically addressing sources of error and uncertainty, the framework ensures that critical safety insights are extracted reliably from unstructured reports, supporting proactive risk management and informed decision-making in industrial environments [15][11][7][2].

7 Applications and Broader Impact

7.1 Deployment in Mining Operations

The deployment of advanced question-answering (QA) systems such as SafetyBERT in mining operations addresses the critical need for extracting actionable safety insights from the vast corpus of unstructured accident reports generated in the industry. Mining environments are characterized by high-risk activities and complex operational contexts, making timely and accurate identification of hazards essential for both regulatory compliance and the prevention of incidents. Traditional approaches relying on manual review of incident narratives are not only labor-intensive but also prone to inconsistencies and delays, especially as the volume of data grows [2]. Integrating a domain-specific language model like SafetyBERT into mining operations enables automated extraction and categorization of safety hazards, management measures, and contributing factors from free-text accident reports. The model is trained to recognize nuanced terminology and context-specific patterns unique to mining, which general-purpose models often fail to capture effectively [51]. For instance, the answer selection mechanism within the QA framework is tailored to match hazard descriptions with appropriate management responses, facilitating the intelligent generation of actionable recommendations for safety management [11]. This approach leverages both positive and negative sample pairs during training, enhancing the model's ability to discern relevant from irrelevant information in complex narratives. A significant challenge in deploying such systems is the scarcity of annotated domain-specific data, as manual labeling of mining safety reports is resource-intensive and requires expert knowledge [19][2]. To mitigate this, the framework incorporates active learning strategies, where the model iteratively selects the most informative and uncertain samples for annotation, thereby maximizing the efficiency of the labeling process [19]. This confidence-based sampling not only reduces the annotation burden but also accelerates model adaptation to the evolving linguistic landscape of mining safety documentation. The continual training of pretrained language models (PLMs) on mining-specific corpora has demonstrated substantial improvements in extracting relevant safety information, as evidenced by applications in both mining and related high-risk domains [2]. By fine-tuning on synthetic or real QA pairs derived from mining accident data, the model achieves higher accuracy in identifying causal factors, environmental conditions, and recommended interventions [19]. For example, semi-automated pipelines

utilizing large language models such as GPT-4o have been employed to summarize crash narratives and attribute key factors, further enhancing the granularity and interpretability of extracted insights [14]. From an operational perspective, the deployment of SafetyBERT-based QA systems can be seamlessly integrated with existing data infrastructures in mining organizations. Accident reports, which are often stored in structured databases or as free-text documents, can be ingested and processed in real time, enabling dynamic risk assessment and proactive safety management. The scalability of the approach ensures that as new data becomes available, the model can be continually updated, maintaining its relevance and effectiveness in rapidly changing operational contexts. Furthermore, the adoption of such intelligent systems supports compliance with regulatory requirements by providing transparent and auditable records of hazard identification and mitigation actions. The ability to rapidly surface critical safety insights from unstructured data not only enhances situational awareness for safety managers but also contributes to a culture of continuous improvement in mining operations [2]. The personalized and context-aware responses generated by the model can be tailored to the specific needs of different user groups, from frontline workers to management, ensuring that safety recommendations are both actionable and relevant [22]. The computational demands of deploying transformer-based models in industrial settings are non-trivial, yet advances in model efficiency and the use of transfer learning have made it feasible to implement such solutions even with limited computational resources [7]. Transfer learning enables the adaptation of large pretrained models to the mining domain using relatively small, high-quality datasets, reducing both training time and the need for extensive labeled data [31]. This is particularly advantageous in mining, where domain-specific datasets are often much smaller than those available in more general domains. In summary, the deployment of SafetyBERT and similar domain-adapted QA systems in mining operations represents a transformative step toward data-driven safety management. By automating the extraction of critical insights from unstructured reports, optimizing annotation workflows through active learning, and ensuring scalability and adaptability, these systems provide a robust foundation for enhancing safety outcomes in high-stakes industrial environments [51][11][19][2][14][22].

7.2 Transferability to Other High-Risk Domains

7.2.1 Construction and Industrial Safety

Construction and industrial safety represent high-risk domains where the consequences of accidents can be severe, both in terms of human life and economic impact. The application of advanced natural language processing (NLP) techniques, particularly domain-adapted language models, has shown significant promise in extracting actionable safety insights from unstructured accident reports and incident narratives. In these sectors, the diversity and complexity of incident descriptions, coupled with the use of specialized terminology, pose unique challenges for generic language models, which often lack the necessary contextual understanding to accurately interpret and classify safety-critical information [2][18]. The transferability of specialized question-answering (QA) frameworks, such as those developed for mining safety, to construction and industrial safety is facilitated by the shared characteristics of high-risk environments. Both domains generate large volumes of unstructured textual data, including accident reports, near-miss descriptions, and safety audits. Leveraging models like SafetyBERT, which are pre-trained or continually adapted on domain-specific corpora, enables more precise extraction of hazards, contributing factors, and preventive recommendations from these texts [2][37]. However, the process of adapting general-purpose models such as BERT to construction and industrial safety is not trivial. Retraining or continual pre-training on domain-specific data is computationally intensive and requires access to substantial annotated datasets, which are often scarce due to the resource-intensive nature of manual data labeling [2][18]. To address these limitations, active learning strategies have been proposed, wherein the model iteratively selects the most informative or uncertain samples for annotation, thereby maximizing performance gains with minimal labeling effort. This approach is particularly advantageous in construction and industrial safety, where expert annotation is costly and time-consuming. By integrating confidence-based sampling with a specialized language model, the framework can efficiently prioritize cases that are likely to improve the model's understanding of rare or complex incident types [39][2]. The use of aspect-based classification further enhances the model's ability to focus on relevant narrative segments, improving the identification of hazards such as scaffolding failures or equipment malfunctions [39]. The effectiveness of domain-specific models in construction and industrial safety has been demonstrated in several studies. For

instance, continual training of pre-trained language models (PLMs) on construction and occupational health datasets has led to improved performance in incident classification and risk factor extraction tasks. In the mining sector, similar approaches have yielded robust models capable of handling the nuanced language and context-specific hazards present in accident reports. The adaptability of these methods to construction and industrial safety is supported by the structural similarities in the data and the types of safety challenges encountered. Moreover, the integration of numerical variables with textual descriptions, as explored in recent research, allows for more comprehensive risk assessments by combining structured and unstructured data sources. This multimodal approach is particularly relevant in industrial settings, where sensor data, inspection logs, and maintenance records can be linked with narrative reports to provide a holistic view of safety performance. The broader impact of deploying such intelligent QA systems in construction and industrial safety extends beyond improved incident analysis. These tools can support proactive safety management by enabling real-time monitoring of safety trends, early detection of emerging hazards, and automated generation of targeted safety recommendations. The scalability of the framework, underpinned by active learning and domain adaptation, ensures that it can be effectively applied across diverse industrial contexts, from large-scale construction projects to manufacturing plants [2][18]. The authors of Fan et al. indicate that leveraging advanced language models for predictive analytics in safety-critical domains can generate insightful predictions even in the presence of data limitations, further underscoring the value of these approaches for construction and industrial safety applications. In summary, the transfer of advanced NLP-based QA frameworks from mining to construction and industrial safety is both feasible and beneficial, provided that domain adaptation and efficient annotation strategies are employed. The combination of specialized language models, active learning, and multimodal data integration offers a scalable solution for extracting critical safety insights and supporting high-stakes decision-making in these high-risk environments [16][39][2].

7.2.2 Healthcare Incident Analysis

Healthcare incident analysis presents a complex challenge due to the highly specialized vocabulary, intricate relationships between clinical events, and the critical importance of accurate information extraction for patient safety. The transferability of advanced NLP frameworks, such as those developed for industrial safety, to healthcare is both promising and nuanced. One of the primary obstacles in this context is the limitation of general-purpose language models, which often struggle to represent domain-specific terminology effectively. For instance, standard BERT models tend to fragment medical terms into multiple subword tokens, reducing their semantic coherence and potentially impairing downstream tasks such as named entity recognition or relation extraction. In contrast, models like PubMedBERT, which incorporate biomedical terms directly into their vocabulary, demonstrate improved handling of specialized language, as evidenced by their ability to represent terms like "naloxone" and "acetyltransferase" as single tokens rather than fragmented subwords **0210406**. This adaptation leads to more accurate and contextually relevant embeddings, which are crucial for extracting actionable insights from unstructured clinical narratives. The effectiveness of domain-adapted models is further supported by comparative studies showing that in-domain pretraining on scientific or biomedical corpora yields measurable improvements in task performance. Beltagy et al. [55] report that SCIBERT, pretrained on scientific texts, achieves higher F1 scores on biomedical and computer science tasks compared to general BERT models, suggesting that exposure to domain-specific language during pretraining is a key factor in enhancing model utility for healthcare incident analysis. Similarly, fine-tuned models such as BlueBERT and BioBERT, which leverage large-scale biomedical datasets, outperform generic architectures in named entity recognition tasks, as demonstrated by their superior F1 scores across various entity types [29]. These findings underscore the necessity of tailoring language models to the unique linguistic characteristics of healthcare data. Manual annotation of healthcare incident reports is resource-intensive, given the need for expert knowledge and the sensitive nature of the data. Active learning (AL) frameworks offer a pragmatic solution by iteratively selecting the most informative samples for annotation, thereby maximizing model improvement while minimizing labeling effort. Zhang et al. [26] highlight that AL is particularly well-suited for tasks where labeled data is scarce or expensive to obtain, which is often the case in healthcare. However, the complexity of healthcare incident analysis, which may involve multi-label classification, temporal reasoning, and the extraction of causal relationships, requires careful adaptation of AL strategies. The integration of confidence-based sampling, as proposed in advanced safety analysis frameworks, can be

leveraged to prioritize ambiguous or uncertain cases for expert review, accelerating the development of robust incident analysis systems. The comparative analysis of embedding techniques also reveals important considerations for healthcare applications. Sentence-level embeddings generated by models such as SBERT capture contextual and relational information more effectively than traditional word embeddings like Word2Vec or GloVe, which require additional pooling strategies to aggregate word-level representations [1]. This property is particularly advantageous in healthcare, where the meaning of clinical events often depends on the interplay between multiple entities and temporal sequences within a narrative. Smetana et al. [4] further note that state-of-the-art embedding models, such as OpenAI's Ada, have demonstrated strong performance in clustering and analyzing safety-related incidents, suggesting their potential applicability to healthcare incident clustering and trend analysis. Domain-specific fine-tuning has been shown to yield substantial gains in classification accuracy for safety-related tasks, including those in healthcare. Danish and Chatterjee [2] provide empirical evidence that models adapted to the occupational safety domain outperform general-purpose models in tasks such as compliance assessment and incident reporting analysis. This observation aligns with the broader trend in NLP, where specialized models consistently surpass their generic counterparts in high-stakes, domain-specific applications. Despite these advances, challenges remain in scaling such systems for real-world healthcare environments. The annotation bottleneck, the need for continual adaptation to evolving medical knowledge, and the integration of human-in-the-loop frameworks are ongoing areas of research. The work of Yugu et al. **0210406** suggests that even sophisticated pretraining strategies, such as adversarial pretraining, may not always yield improvements and can sometimes degrade performance, highlighting the importance of empirical validation in each new domain. The transfer of advanced NLP techniques from industrial safety to healthcare incident analysis is thus characterized by both significant opportunities and domain-specific challenges. The evidence indicates that with appropriate adaptation, particularly through domain-specific pretraining, active learning, and the use of sentence-level embeddings, these frameworks can substantially enhance the extraction of critical safety insights from unstructured healthcare data, supporting more effective incident analysis and ultimately contributing to improved patient safety [1][4][29][55]**0210406**[2][26].

7.2.3 Aviation and Transportation Safety

Aviation and transportation safety present unique challenges for natural language processing (NLP) systems due to the complexity, heterogeneity, and high-stakes nature of incident data. The transferability of advanced question-answering (QA) frameworks, such as those developed for mining safety, to these domains is both promising and technically demanding. The integration of pre-trained language models, particularly those fine-tuned on domain-specific corpora, has demonstrated substantial improvements in extracting actionable insights from unstructured safety reports and accident narratives [1]. In aviation, for instance, the volume and diversity of incident reports necessitate robust models capable of discerning subtle contextual cues and rare event patterns, which are often underrepresented in general-purpose datasets. Continual training of pre-trained language models (PLMs) on aviation and transportation-specific data has emerged as an effective strategy to address domain adaptation challenges. This approach leverages the foundational linguistic knowledge encoded in models like BERT, while incrementally incorporating specialized vocabulary and context from aviation safety documentation [2]. The authors of **0210406** indicate that pretraining on in-domain text, such as aviation incident reports, corrects errors that persist when relying solely on general-domain pretraining, underscoring the necessity of domain-specific adaptation for high-fidelity information extraction. The application of active learning within this context is particularly advantageous. Manual annotation of aviation safety data is resource-intensive, given the technical complexity and the need for expert knowledge. By employing confidence-based sampling strategies, the annotation process can be optimized, focusing human effort on the most informative and uncertain cases. This not only accelerates the creation of high-quality labeled datasets but also enhances model performance with minimal annotation overhead [2]. Such strategies are crucial in aviation, where the rapid identification of emerging risk factors can have immediate operational and regulatory implications. Semantic-rich sentence embeddings, generated by models like SBERT, further enhance the analytical capabilities of safety QA systems in transportation domains. These embeddings facilitate the clustering and retrieval of incident narratives with similar causal structures or risk factors, enabling more systematic analysis of safety trends and the identification of latent hazards [1]. The versatility of this methodology allows for seamless adaptation to various transportation modalities, including rail, maritime, and road traffic, each characterized

by distinct operational contexts and safety concerns. Recent advances in visualization techniques, such as aspect-level risk factor association diagrams, provide interpretable representations of the relationships between contributing factors and safety outcomes in transportation incidents [14]. These visualizations support domain experts in tracing the propagation of risk across complex event chains, thereby informing targeted interventions and policy decisions. The integration of such interpretability tools within QA frameworks enhances their practical utility in high-risk environments. The challenges posed by large, noisy contexts in transportation safety documentation, where relevant information is often embedded within extensive, heterogeneous narratives, necessitate efficient answer selection and context filtering mechanisms [19]. Accurate identification of pertinent text snippets is essential for reliable QA performance, especially when dealing with regulatory compliance, accident investigation, and real-time operational support. The comparative evaluation of extractive and generative QA models in transportation safety settings reveals complementary strengths. Extractive models, leveraging cosine similarity retrievers, excel at pinpointing precise answers within structured datasets, while generative large language models offer conversational interfaces that can synthesize information from multiple sources, grounded in real-world incident data [18]. This dual approach supports both granular fact retrieval and broader situational awareness, which are critical for safety management in aviation and transportation. The cumulative evidence from these studies demonstrates that the methodological innovations developed for mining safety, such as domain-adaptive pretraining, active learning, and semantic clustering, are readily transferable to aviation and transportation safety. These techniques collectively enable scalable, interpretable, and high-precision analysis of unstructured safety data, supporting proactive risk mitigation and continuous improvement in high-stakes operational environments [1][2][19]**0210406**[14][18].

7.3 Enhancing Proactive Hazard Mitigation

Enhancing proactive hazard mitigation in mining safety environments requires the integration of advanced natural language processing (NLP) techniques capable of extracting actionable insights from vast, unstructured accident reports. Traditional approaches relying on general-purpose language models often fall short in specialized domains due to their limited understanding of domain-specific terminology and context, which can result in suboptimal hazard identification and delayed mitigation actions [46][50]. The deployment of specialized models, such as SafetyBERT, addresses these limitations by leveraging domain-adapted pre-training strategies that expose the model to mining-specific language and safety concepts, thereby improving its ability to recognize subtle indicators of risk and emerging hazards [46][50][8]. The process of proactive hazard mitigation is fundamentally dependent on the timely and accurate extraction of critical information from incident narratives. General-domain models, even when fine-tuned, may not sufficiently capture the nuanced patterns and terminology unique to mining safety, leading to missed opportunities for early intervention [50][8]. By contrast, models pre-trained or further adapted on specialized corpora demonstrate superior performance in identifying relevant safety events, causal factors, and near-miss scenarios, which are essential for anticipating and preventing future incidents [46]. The work of Peng et al. [50] highlights that continuous pre-training on domain-specific text yields more consistent improvements across tasks, even when the vocabulary is not explicitly specialized, underscoring the value of contextual adaptation for proactive risk management. Manual annotation of mining safety data is resource-intensive and often impractical at scale, especially given the volume and heterogeneity of accident reports generated in industrial settings [43]. To address this challenge, the integration of active learning frameworks into the model development pipeline enables more efficient utilization of limited labeled data. Active learning strategies, particularly those employing confidence-based sampling, prioritize the selection of informative and uncertain samples for annotation, thereby accelerating model improvement while minimizing annotation costs [15][2]. Zhu et al. [15] propose a multitask active learning framework that jointly optimizes intent detection and slot filling, demonstrating that leveraging relational information between tasks can further enhance sample selection and model performance in natural language understanding applications. The combination of a specialized language model with an intelligent active learning strategy creates a scalable and adaptive tool for mining safety analysis. This synergy allows for continuous refinement of the model as new data becomes available, ensuring that the system remains responsive to evolving operational risks and regulatory requirements [43][15]. Danish et al. [2] emphasize the importance of addressing source distribution imbalance during training, ensuring that underrepresented but critical safety scenarios are adequately captured and reflected in the model's predictions. Such balanced rep-

resentation is crucial for comprehensive hazard mitigation, as rare but high-impact events must not be overlooked. Furthermore, the application of advanced question-answering (QA) systems in mining safety supports the automatic generation of safety hazard management measures, facilitating rapid dissemination of best practices and corrective actions [43][11]. Generation-based QA approaches, which synthesize new answers from a given corpus rather than relying solely on predefined question-answer pairs, are particularly valuable in dynamic environments where novel hazards may emerge [43]. The ability to automatically process and interpret massive volumes of safety data within a short timeframe not only enhances the efficiency of hazard mitigation efforts but also promotes the widespread adoption of intelligent safety management systems [11]. The integration of BERT-based retrieval and passage selection modules further refines the system's ability to pinpoint the most relevant information within lengthy and complex accident reports. Frermann [27] observes that such retrieval systems effectively narrow down the context to a small subset of highly relevant passages, enabling more precise extraction of hazard-related insights and supporting targeted mitigation strategies. By leveraging these advancements, mining organizations can transition from reactive to proactive safety management paradigms. The continuous, automated extraction and analysis of safety-critical information empower decision-makers to anticipate risks, implement preventive measures, and ultimately reduce the incidence and severity of workplace accidents [43][15][2][11]. This approach not only enhances operational safety but also contributes to the broader goal of sustainable and responsible resource extraction.

7.4 Future Directions in Intelligent Safety Management

The trajectory of intelligent safety management is shaped by the integration of advanced natural language processing (NLP) models, domain-specific adaptation, and scalable learning strategies. As industrial environments generate vast quantities of unstructured safety data, the need for systems that can autonomously extract, interpret, and act upon critical safety information becomes increasingly urgent. The application of specialized language models, such as SafetyBERT, demonstrates the potential for tailored NLP solutions to outperform general-purpose models in extracting nuanced risk factors from accident narratives, particularly when paired with domain-specific prompting and structured output formats that highlight salient risk categories. This approach not only enhances interpretability but also supports the identification of systemic safety issues through multi-level attribution analysis, enabling both granular and aggregate insights into risk factors [14]. A significant challenge in deploying such systems is the resource-intensive nature of manual data annotation, which is often a bottleneck in developing robust, domain-adapted models. Active learning (AL) emerges as a promising solution, allowing for the efficient selection of informative samples that maximize model improvement while minimizing annotation effort [9]. The authors of [15] indicate that multitask active learning frameworks, which jointly optimize for multiple related tasks, can further enhance sample efficiency and model generalization in natural language understanding (NLU) applications. In the context of safety management, this suggests that future systems could leverage joint modeling of incident detection and severity factor extraction, selecting data points that benefit both tasks simultaneously. The automation of information extraction from accident reports is another area poised for advancement. While current studies often rely on direct interaction with large language model (LLM) interfaces, the transition to fully automated pipelines via API integration will enable real-time, scalable safety analysis [12]. This shift is particularly relevant for high-stakes environments where timely identification of hazards can prevent incidents and save lives. Moreover, the development of knowledge-based question answering systems, as demonstrated in [11], illustrates how answer selection models can be harnessed to generate actionable safety management measures, further closing the loop between data analysis and operational decision-making. Interpretability remains a central concern, especially as LLMs are increasingly deployed in critical domains. Token-level attribution methods, which identify the specific factors influencing model predictions, provide transparency and support trust in automated safety assessments. By systematically analyzing attribution patterns across incidents, organizations can uncover latent safety trends and prioritize interventions. The structured summarization capabilities described in [14] exemplify how domain-specific models can deliver concise, actionable insights tailored to the needs of safety professionals. The broader impact of these advancements extends beyond mining safety to other high-risk industries, such as finance and healthcare, where the volume and complexity of unstructured data similarly challenge human analysts. The adoption of deep learning and NLP-driven question answering systems in these domains underscores the generalizability of the proposed approaches and highlights the potential for cross-domain innovation [24]. As Bayesian optimization and other hyper-

parameter tuning techniques continue to improve model performance in question answering tasks, as shown in [7], the effectiveness of information extraction systems will further increase, supporting more accurate and comprehensive safety analyses. Future directions in intelligent safety management will likely involve the convergence of several key trends: the refinement of domain-specific language models, the integration of active and multitask learning strategies, the automation of data processing pipelines, and the enhancement of interpretability through attribution analysis. The combination of knowledge graph-based and deep learning-based question answering systems, as discussed by Ankush Agarwal et al. [38], offers a pathway to overcoming the limitations of individual approaches, enabling more robust and context-aware safety intelligence. As these technologies mature, the intelligent management of safety information will become increasingly proactive, data-driven, and adaptive to the evolving complexities of industrial operations.

8 Discussion

8.1 Opportunities and Limitations

The development of a domain-specific question-answering system for mining safety, such as SafetyBERT, presents significant opportunities for advancing safety analysis in high-risk industrial contexts. By leveraging advanced natural language processing techniques, these systems can extract actionable insights from unstructured accident reports, which are often rich in detail but challenging to process manually. The use of specialized language models, particularly those pre-trained on domain-relevant corpora, has demonstrated clear advantages in extracting nuanced information that general-purpose models may overlook. For instance, models like MechBERT, which are tailored to specific scientific domains, have shown superior performance in extractive tasks due to their exposure to domain-specific terminology and context during pretraining [7]. Similarly, CancerBERT's adaptation to the clinical domain resulted in improved extraction of complex phenotypes from medical texts, underscoring the value of domain adaptation for specialized tasks [29]. The integration of active learning strategies further enhances the efficiency and scalability of such systems. Active learning addresses the resource-intensive nature of manual data annotation by prioritizing the selection of the most informative samples for labeling, thereby reducing the overall annotation burden while maximizing model improvement. This is particularly relevant in industrial safety, where labeled data is scarce and costly to obtain. The batch mode active learning paradigm, necessitated by the computational demands of deep neural networks, allows for practical model updates without retraining from scratch for each new sample [13]. Moreover, the combination of knowledge distillation and active learning has been shown to reduce model complexity and labeling costs, making it feasible to deploy high-performing models in environments with limited resources [9]. Despite these opportunities, several limitations persist. One of the primary challenges is the calibration of model confidence. Large language models, including those adapted for specific domains, often lack well-calibrated mechanisms for assessing the correctness of their outputs. This can lead to overconfident predictions, particularly in high-stakes scenarios where erroneous outputs may have severe consequences. Addressing this limitation requires the development of improved calibration techniques and the integration of external tools or knowledge bases to better estimate and communicate uncertainty [20]. Additionally, the black-box nature of many advanced NLP models poses challenges for transparency and interpretability, which are critical for building trust in safety-critical applications. Post-hoc explanation methods provide some insight, but they may not fully address the need for detailed, context-aware explanations that can be scrutinized by domain experts [14]. Another limitation arises from the transferability of general models to specialized domains. While models like BERT have achieved state-of-the-art results across a range of NLP tasks, their performance can degrade when applied to domains with unique linguistic characteristics or specialized vocabularies [24]. The necessity to generate domain-specific vocabularies and adapt model architectures for multi-sentence or context-rich inputs further complicates deployment **0210406**[24]. Furthermore, the effectiveness of additional pretraining diminishes when ample labeled data is available, suggesting diminishing returns for continued model adaptation in well-resourced domains [17]. The process of verifying the quality and accuracy of model outputs also introduces challenges. Automated verification using large language models or retrieval-augmented generation (RAG) frameworks can help ensure the stability and uniqueness of answers, but these methods are not infallible and may require human oversight for critical decisions [56]. The reliance on statistical analysis and comparative

studies to validate model performance highlights the ongoing need for rigorous evaluation methodologies, particularly as models are deployed in dynamic, real-world settings [36]. Opportunities for future research include the development of more robust uncertainty estimation methods, improved interpretability frameworks, and the exploration of hybrid approaches that combine domain-specific pretraining with active learning and external knowledge integration. The authors of indicate that learning policies to decide when expert intervention is necessary can further optimize the annotation process and enhance model reliability. As the field progresses, addressing these limitations will be essential for realizing the full potential of domain-specific question-answering systems in mining safety and other high-stakes industrial applications [26][36].

8.2 Ethical and Practical Considerations

The development and deployment of a domain-specific question-answering system for mining safety, particularly one leveraging advanced NLP models such as SafetyBERT, necessitate careful consideration of both ethical and practical dimensions. One of the foremost ethical concerns is the potential for bias in the training data and model outputs. Mining accident reports often originate from diverse sources, each with its own reporting standards and terminologies. If the model is trained predominantly on data from a subset of sources, there is a risk that underrepresented perspectives or specialized industry terminology may be inadequately captured, leading to skewed or incomplete safety insights. Danish et al. [2] highlight the importance of ensuring that sequences from underrepresented sources contribute proportionally more during training, which helps counteract source distribution imbalance and ensures that specialized terminology is adequately represented. Another ethical aspect involves the transparency and interpretability of the model's decisions. In high-stakes environments such as mining safety, stakeholders must be able to understand and trust the rationale behind the system's recommendations. Deep learning models, especially those based on transformer architectures, are often criticized for their black-box nature. The authors of [11] indicate that while deep learning enables comprehensive semantic analysis and effective question-answering, it can also obscure the reasoning process, making it challenging for users to validate or contest the system's outputs. This opacity can have significant implications if the system is used to inform critical safety decisions. The resource-intensive nature of manual data annotation presents both ethical and practical challenges. Manual annotation is not only costly and time-consuming but also exposes annotators to potentially distressing content, such as detailed descriptions of accidents. The active learning approach proposed by Afzal et al. and further discussed in aims to minimize the labor cost of manual annotation by intelligently selecting the most informative samples for labeling. This strategy not only improves efficiency but also reduces the exposure of human annotators to sensitive or traumatic information, addressing a key ethical concern. From a practical standpoint, the integration of domain knowledge into the model is essential for achieving high performance in specialized fields. Afzal et al. [5] demonstrate that embedding techniques incorporating both domain-specific and domain-independent knowledge can enhance the reusability and adaptability of the model across various clinical or safety-related document classification tasks. However, the initial seed dataset, often generated using rule-based approaches, plays a critical role in determining the quality of the active learning process. If the seed data is not representative or contains systematic biases, these issues may propagate through subsequent iterations, potentially compromising the reliability of the system. The challenge of generalizing models trained on one domain to another is well-documented. Yugu et al. **0210406** observe that BERT models pretrained on clinical notes perform poorly on biomedical tasks, and vice versa, underscoring the necessity of creating separate benchmarks and models for distinct domains. This observation is echoed by Kanakarajan et al. [3], who note that models such as ClinicalBERT, SciBERT, and PubMedBERT are pretrained on different types of domain-specific data to optimize performance. For mining safety, this means that a dedicated model like SafetyBERT must be carefully tailored and validated on mining-specific corpora to ensure its effectiveness and ethical soundness. The use of active learning introduces additional practical considerations. Boreshban et al. [9] discuss that active learning strategies rely on criteria such as uncertainty measures to select unlabeled data for annotation. While this can accelerate model improvement, it also raises questions about the representativeness of the selected samples and the potential for reinforcing existing biases if the selection criteria are not carefully designed. Furthermore, the iterative nature of active learning requires ongoing monitoring to ensure that the model does not drift away from its intended purpose or inadvertently prioritize certain types of incidents over others. Data privacy and confidentiality are also critical ethical considerations, especially when

handling sensitive accident reports. The aggregation and analysis of such data must comply with relevant legal and regulatory frameworks to protect the identities and rights of individuals involved in reported incidents. Ricketts et al. [18] suggest that expanding the knowledge base to include diverse sources such as safety assessment reports and maintenance data can enhance the system's utility, but this also increases the complexity of managing data privacy and access controls. Finally, the scalability and adaptability of the proposed framework are essential for practical deployment in industrial settings. The rapid evolution of transformer-based models, as noted by Chen and Zulkernine [43], has led to significant improvements in the performance and efficiency of question-answering systems. However, the adoption of such models in safety-critical domains requires rigorous validation, continuous monitoring, and mechanisms for human oversight to ensure that the system remains aligned with ethical standards and operational requirements. In summary, the ethical and practical considerations in developing a domain-specific question-answering system for mining safety encompass issues of bias, transparency, data privacy, annotation burden, domain adaptation, and system scalability. Addressing these challenges is crucial for building a trustworthy and effective tool that can support high-stakes decision-making in industrial environments [5][11][2][43]**0210406**[18][3][9].

8.3 Potential for Industry-Wide Transformation

The integration of advanced NLP techniques, such as SafetyBERT, into mining safety analysis holds significant promise for transforming safety management practices across the industry. By leveraging domain-specific language models tailored to the unique linguistic patterns and terminologies found in accident reports, organizations can systematically extract actionable insights that were previously buried in unstructured text. This capability enables a shift from reactive to proactive safety management, as critical hazards and near-miss patterns can be identified and addressed before they escalate [5][12]. The challenge of applying general-purpose models to specialized domains like mining safety is well documented. Generic models often struggle to capture the nuanced semantics and context-specific vocabulary present in technical narratives, leading to suboptimal performance in extracting relevant information [40][12]. The development and deployment of specialized models, such as SafetyBERT, directly address this limitation by incorporating domain knowledge during pre-training and fine-tuning phases, resulting in improved accuracy and relevance of extracted insights [5][39]. Manual annotation of safety data is a resource-intensive process, often requiring expert knowledge and significant time investment. This bottleneck has historically limited the scalability of data-driven safety analysis. The adoption of active learning strategies, particularly those that utilize confidence-based sampling, offers a practical solution by prioritizing the annotation of the most informative samples [9][13]. This approach reduces the overall labeling effort while maximizing model improvement, making it feasible to continuously refine the system as new data becomes available [9]. The iterative nature of active learning ensures that the model remains adaptive to evolving operational contexts and emerging risks. The potential for industry-wide transformation is further amplified by the scalability and adaptability of the proposed framework. Once established, the combination of a specialized language model and an intelligent sampling strategy can be extended to other high-risk industrial sectors, such as construction, oil and gas, or chemical manufacturing, where unstructured safety narratives are prevalent [18][12]. The modularity of the approach allows for the incorporation of additional data sources, such as maintenance logs or hazard analyses, thereby enriching the knowledge base and supporting more comprehensive risk assessments. Moreover, the automation of safety insight extraction can lead to substantial resource savings. By minimizing redundant hazard identification efforts and enabling knowledge transfer across projects and teams, organizations can allocate their safety engineering resources more efficiently [18]. The improved accessibility and timeliness of safety intelligence support informed decision-making at both operational and strategic levels, potentially reducing incident rates and improving overall workplace safety [39][12]. The use of transformer-based architectures, such as BERT and its derivatives, has already demonstrated superior performance in various NLP tasks, including question answering and text classification, particularly when adapted to domain-specific requirements [28][31][18]. The ability to fine-tune these models on specialized corpora, despite challenges related to input length and data sparsity, further enhances their applicability to industrial safety contexts [40][18]. The integration of active learning not only addresses the annotation bottleneck but also ensures that the models remain robust and generalizable as new types of incidents and hazards are encountered [13][9]. In summary, the deployment of a domain-specific question-answering system powered by advanced NLP and active learning methodologies has the capacity to fundamentally reshape safety analysis in mining and related

industries. The resulting improvements in efficiency, accuracy, and scalability position such systems as key enablers for data-driven safety management in high-stakes environments [5][9][18][12].

9 Conclusion

The advancement of domain-specific question-answering systems tailored for mining safety represents a pivotal development in the pursuit of enhanced occupational safety within high-risk industrial environments. By harnessing the power of specialized natural language processing models, such as SafetyBERT, these systems effectively bridge the gap between unstructured accident narratives and actionable safety insights. The integration of domain-adapted pre-training ensures that the unique linguistic characteristics and technical terminology inherent to mining safety are accurately captured, enabling more precise extraction and interpretation of critical information that general-purpose models often overlook.

Addressing the significant challenge of limited annotated data, the incorporation of active learning frameworks optimizes the annotation process by prioritizing the most informative and uncertain samples for expert review. This strategy not only reduces the manual labeling burden but also accelerates model refinement, ensuring that the system remains responsive to evolving safety concerns and diverse incident types. The iterative training and feedback mechanisms embedded within this approach facilitate continuous improvement, allowing the model to adapt dynamically to new data and maintain high performance despite the complexity and variability of safety reports.

Furthermore, the deployment of such intelligent QA systems within mining operations offers transformative potential for proactive hazard identification and risk mitigation. By automating the extraction of nuanced safety factors and management measures from vast textual datasets, organizations can transition from reactive to anticipatory safety management paradigms. This shift enhances operational resilience, supports regulatory compliance, and ultimately contributes to the reduction of workplace accidents and associated human and economic costs.

The scalability and adaptability of the framework extend its applicability beyond mining, with promising transferability to other high-risk sectors such as construction, industrial manufacturing, healthcare, and transportation safety. The shared challenges of unstructured data, domain-specific language, and annotation scarcity in these fields can be effectively addressed through the combination of domain adaptation, active learning, and advanced transformer architectures. This cross-industry relevance underscores the broader impact of the methodologies developed.

Despite these advances, challenges remain in ensuring model transparency, mitigating biases inherent in training data, and maintaining data privacy and ethical standards. Continued efforts to improve uncertainty estimation, interpretability, and human-in-the-loop collaboration are essential to build trust and facilitate the responsible adoption of these technologies in safety-critical contexts. Moreover, ongoing research into optimizing model efficiency and annotation workflows will further enhance the practicality of deploying such systems at scale.

In summary, the integration of domain-specific language models with intelligent annotation strategies marks a significant stride toward more effective, scalable, and interpretable safety analysis tools. These innovations hold the promise of fundamentally transforming safety management practices, enabling data-driven decision-making that safeguards workers and assets in mining and other high-stakes industrial environments. The continued evolution and refinement of these approaches will be instrumental in advancing occupational safety and fostering a culture of proactive risk mitigation across diverse sectors.

References

[1] U. Author, *Contents*, Oct. 2023.

[2] A. A. K. Danish and S. Chatterjee, "Bridging the safety -specific language model gap: Domain -adaptive pretraining of transformer -based models across several industrial sectors for occupational safety applications", Oct. 2023.

[3] K. R. Kanakarajan, B. Kundumani, and M. Sankarasubbu, "Bioelectra: Pretrained biomedical text encoder using discriminators", Jun. 2021, pp. 143–154. [Online]. Available: https://github.com/kamalkraj/BioELECTRA.

[4] M. Smetana, L. S. de Salles, I. Sukharev, and L. Khazanovich, "Highway construction safety analysis using large language models", *Applied Sciences*, vol. 14, p. 1352, Feb. 2024. DOI: 10.3390/app14041352. [Online]. Available: https://doi.org/10.3390/app14041352.

[5] M. Afzal, J. Hussain, A. Abbas, and M. Hussain, "Multi-class clinical text annotation and classification using bert-based active learning", *Expert Systems with Applications*, Mar. 2022. [Online]. Available: https://ssrn.com/abstract=4081033.

[6] S. Kierszbaum and L. Lapasset, "Applying distilled bert for question answering on asrs reports", Oct. 2023.

[7] PankajKumar, SaurabhKabra, and J. M.Cole, "Mechbert: Language models for extracting chemical and property relationships about mechanical stress and strain", *J. Chem. Inf. Model.*, vol. 65, pp. 1873–1888, Jan. 2025. DOI: 10.1021/acs.jcim.4c00857.

[8] Y. H. GU, X. PIAO, H. YIN, D. JIN, R. ZHENG, and S. J. YOO, "Domain-specific language model pre-training for korean tax law classification", Apr. 2022. DOI: 10.1109/ACCESS.2022.3164098.

[9] Y. Boreshban, S. M. Mirbostani, G. Ghassem-Sani, S. A. Mirroshandel, and S. Amiriparian, "Improving question answering performance using knowledge distillation and active learning", *Engineering Applications of Artificial Intelligence*, vol. 123, p. 106 137, Mar. 2023. DOI: 10.1016/j.engappai.2023.106137. [Online]. Available: https://github.com/mirbostani/QA-KD-AL.

[10] K. Margatina, L. Barrault, and N. Aletras, "On the importance of effectively adapting pretrained language models for active learning", May 2022, pp. 825–836.

[11] D. Tiana, M. Li, Q. Ren, X. Zhang, S. Han, and Y. Shen, "Intelligent question answering method for construction safety hazard knowledge based on deep semantic mining", *Automation in Construction*, vol. 145, p. 104 670, Nov. 2023. DOI: 10.1016/j.autcon.2022.104670. [Online]. Available: https://doi.org/10.1016/j.autcon.2022.104670.

[12] M. Mumtarin, M. S. Chowdhury, and J. Wood, *Large language models in analyzing crash narratives - a comparative study of chatgpt, bard and gpt-4*, Aug. 2023. [Online]. Available: arXiv:2308.13563v1.

[13] L. Ein-Dor, A. Halfon, A. Gera, *et al.*, "Active learning for bert: An empirical study", Nov. 2020, pp. 7949–7962.

[14] H. Zhen and J. J. Yang, "Crash sage: A large language model -centered framework for contextual and interpretable traffic crash analysis", May 2025.

[15] H. Zhu, W. Ye, S. Luo, and X. Zhang, "A multitask active learning framework for natural language understanding", Dec. 2020, pp. 4900–4914.

[16] Z. Fan, P. Wang, Y. Zhao, *et al.*, "Learning traffic crashes as language: Datasets, benchmarks, and what–if causal analyses", Jun. 2024. [Online]. Available: https://arxiv.org/abs/2406.10789.

[17] M. Bosley, M. Jacobs-Harukawa, H. Licht, and A. Hoyle, "Do we still need bert in the age of gpt? comparing the benefits of domain-adaptation and in-context-learning approaches to using llms for political science research", Apr. 2023.

[18] J. Ricketts, W. Guo, J. Pelham, and D. Barry, "Integrating an incident dataset with a question and answering language model to assist hazard identification: Comparison of an extractive and generative model", *Proc IMechE Part O: J Risk and Reliability*, pp. 1–18, Jul. 2024. DOI: 10.1177/1748006X241272831. [Online]. Available: https://journals.sagepub.com/home/pio.

[19] R. Sarkar, S. Dutta, H. Assem, M. Arcan, and J. McCrae, "Semantic aware answer sentence selection using self-learning based domain adaptation", Aug. 2022. DOI: 10.1145/3534678.3539162. [Online]. Available: https://doi.org/10.1145/3534678.3539162.

[20] M. Yue, "A survey of large language model agents for question answering", Mar. 2025.

[21] N. Shah, H. K. Thakkar, and H. Mewada, *On the analysis of a bert-based domain-specific question answering models for indian legal system*, Jun. 2024.

[22] M. Sammoudi, A. Habaybeh, H. I. Ashqar, and M. Elhenawy, "Question-answering (qa) model for a personalized learning assistant for arabic language",

[23] W. Yang, Y. Xie, A. Lin, *et al.*, "End-to-end open-domain question answering with bertserini", Jun. 2019, pp. 72–77.

[24] B. Yuan, *Finbert-qa: Financial question answering with pre-trained bert language models*, Jul. 2020. DOI: 10.48550/arXiv.2505.00725. [Online]. Available: https://arxiv.org/abs/2505.00725v1.

[25] S. Prabhu, M. Mohamed, and H. Misra, "Multi-class text classification using bert-based active learning", Sep. 2021.

[26] Z. Zhang, E. Strubell, and E. Hovy, "A survey of active learning for natural language processing", Feb. 2023.

[27] L. Frermann, "Extractive narrativeqa with heuristic pre-training", in *Proceedings of the Second Workshop on Machine Reading for Question Answering*, Nov. 2019, pp. 172–182.

[28] X. Qiu, T. Sun, Y. Xu, Y. Shao, N. DAI, and X. HUANG, "Pre-trained models for natural language processing: A survey", *Science China Technological Sciences*, vol. 63, no. 10, pp. 1872–1897, Oct. 2020. DOI: 10.1007/s11431-020-1647-3. [Online]. Available: https://doi.org/10.1007/s11431-020-1647-3.

[29] S. Zhou, N. Wang, L. Wang, H. Liu, and R. Zhang, "Cancerbert: A cancer domain-specific language model for extracting breast cancer phenotypes from electronic health records", *Journal of the American Medical Informatics Association*, pp. 1208–1216, Mar. 2022. DOI: 10.1093/jamia/ocac040.

[30] L. Wang, K. Zheng, L. Qian, and S. Li, "A survey of extractive question answering", Dec. 2022. DOI: 10.1109/HDIS56859.2022.9991478.

[31] R. Jha and V. S. Devi, "Extractive question answering using transformer-based lm", pp. 373–384, Oct. 2023. DOI: 10.1007/978-981-99-1642-9. [Online]. Available: https://www.iisc.ac.in/.

[32] A. Li and J. Wang, "Research on construction site safety q&a system based on bert", *International Journal of Advanced Network, Monitoring and Controls*, vol. 0, no. 4, Apr. 2024. DOI: 10.2478/ijanmc-2024-0039.

[33] O. Khattab, C. Potts, and M. Zaharia, "Relevance-guided supervision for openqa with colbert", pp. 929–944, Sep. 2021. DOI: 10.1162/tacl_a_00405. [Online]. Available: https://doi.org/10.1162/tacl%5C_a%5C_00405.

[34] Y. Liu, M. Ott, N. Goyal, *et al.*, "Roberta: A robustly optimized bert pretraining approach", Jul. 2019. [Online]. Available: https://arxiv.org/abs/1907.11692.

[35] J. Y. Wong, C. P. Lee, K. M. Lim, J. Y. Lim, and J. N. Mogan, *Question answering with language models*, Dec. 2024.

[36] Y. Qi, X. Zhao, S. Khastgir, and X. Huang, *Safety analysis in the era of large language models: A case study of stpa using chatgpt*, Dec. 2023.

[37] S. R. Andrade and H. S. Walsh, "Safeaerobert: Towards a safety -informed aerospace -specific language model", Oct. 2023.

[38] A. Agarwal, R. Gite, S. Laddha, *et al.*, "Knowledge graph – deep learning: A case study in question answering in aviation safety domain", Jun. 2022.

[39] H. A. M. Hassan, E. Marengo, and W. Nutt, "A bert-based model for question answering on construction incident reports", Oct. 2022.

[40] C. E. de Lima Joaquim and T. de Paulo Faleiros, *Bert self-learning approach with limited labels for document classification*, Oct. 2022.

[41] M. Yuan, H.-T. Lin, and J. Boyd-Graber, "Cold-start active learning through self-supervised language modeling", Nov. 2020, pp. 7935–7948.

[42] D. Grießhaber, J. Maucher, and N. T. Vu, "Fine-tuning bert for low-resource natural language understanding via active learning", Dec. 2020, pp. 1158–1171.

[43] Y. Chen and F. Zulkernine, "Bird-qa: A bert -based information retrieval approach to domain specific question answering", in *2021 IEEE International Conference on Big Data (Big Data)*, May 2021, p. 3503. DOI: 10.1109/BigData52589.2021.9671523.

[44] J. Kim, S. Chung, S. Moon, and S. Chi, *Feasibility study of a bert-based question answering chatbot for information retrieval from construction specifications*, Nov. 2022. DOI: 10.1109/IEEM55944.2022.9989625.

[45] S. Ji, T. Zhang, L. Ansari, J. Fu, P. Tiwari, and E. Cambria, "Mentalbert: Publicly available pretrained language models for mental healthcare", *arXiv preprint arXiv:2110.15621v1*, Oct. 2021.

[46] H. E. Boukkouri, O. Ferret, T. Lavergne, and P. Zweigenbaum, "Re-train or train from scratch? comparing pre-training strategies of bert in the medical domain", Jun. 2022, pp. 2626–2633.

[47] V. Siddeshwar, A. Azim, S. Alwidian, and M. Makrehchi, "Towards enhancing aviation safety through advanced incident analysis using large language models", Apr. 2024.

[48] Y. Yang, M. C. S. Uy, and A. Huang, *Finbert: A pretrained language model for financial communications*, Jul. 2020. [Online]. Available: https://github.com/yya518/FinBERT.

[49] R. O. Popov, N. V. Karpenko, and V. V. Gerasimov, "Overview of small language models in practice", Dec. 2025, pp. 164–182.

[50] B. Peng, E. Chersoni, Y.-Y. Hsu, and C.-R. Huang, "Is domain adaptation worth your investment? comparing bert and finbert on financial tasks", Nov. 2021, pp. 37–44.

[51] Z. Liu, D. Huang, K. Huang, Z. Li, and J. Zhao, "Finbert: A pre-trained financial language representation model for financial text mining", *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20) Special Track on AI in FinTech*, Oct. 2020.

[52] K. Pearce, T. Zhan, A. Komanduriy, and J. Zhan, "A comparative study of transformer-based language models on extractive question answering", *arXiv preprint arXiv:2110.03142*, Oct. 2023. [Online]. Available: https://arxiv.org/abs/2110.03142.

[53] W. Zheng, S. Lu, Z. Cai, R. Wang, L. Wang, and L. Yin, "Pal-bert: An improved question answering model", Mar. 2024. DOI: 10.32604/cmes.2023.046692.

[54] Z. Wang, P. Ng, X. Ma, R. Nallapati, and B. Xiang, "Multi-passage bert: A globally normalized bert model for open-domain question answering", Nov. 2019, pp. 5878–5882.

[55] I. Beltagy, K. Lo, and C. Arman, "Scibert: A pretrained language model for scientific text", Sep. 2019. [Online]. Available: https://github.com/allenai/scibert/.

[56] Y. Tan, B. Zheng, B. Zheng, *et al.*, *Chinese safetyqa: A safety short-form factuality benchmark for large language models*, Dec. 2024. [Online]. Available: https://openstellarteam.github.io/ChineseSimpleQA/.