applied sciences

MDPI

Article

Automated Safety Precaution Generation in High-Risk Industries: A Parameter-Efficient Fine-Tuning Approach with Mistral-7B

Hasan Eker [1] and Cihan Bayraktar [2,*]

[1] Property Protection and Safety Division, Occupational Health and Safety Program, Karabuk University, 78400 Karabuk, Türkiye; hasaneker@karabuk.edu.tr
[2] Computer Technologies Department, Karabuk University, 78400 Karabuk, Türkiye
* Correspondence: cihanbayraktar@karabuk.edu.tr; Tel.: +90-506-621-44-41

Abstract

The mining industry faces complex operational hazards that necessitate systematic risk assessments to enable proactive accident prevention. While Large Language Models (LLMs) offer significant potential for the automated generation of safety measures, the limited availability of domain-specific terminology and high-quality labelled safety data (in low-resource environments) hinders their direct application. This study investigates and optimises data augmentation strategies to fine-tune LLMs to generate accurate, context-sensitive safety measures from structured coal mine risk records. The study systematically explored four experimental configurations, leveraging the Mistral-7B-Instruct model in conjunction with Quantised Low-Rank Adaptation (QLoRA) for efficient fine-tuning. These configurations comprised: (i) a baseline without augmentation, (ii) input-side lexical augmentation, (iii) output-side multi-reference augmentation, and (iv) a combined strategy. Performance was measured using BLEU, ROUGE, METEOR, and BERTScore metrics, along with statistical significance testing and qualitative analyses. The results show that, compared to other strategies, the input-side data augmentation strategy performs better. The findings indicate that input-side data augmentation yields significant improvements; this strategy increased the BERTScore (F1) from 0.360 to 0.530 and the BLEU score from 16.02 to 29.50 compared to the baseline model. In contrast, output-side multi-reference augmentation contributed to greater learning uncertainty and a consequent decline in performance. Statistical and qualitative analyses confirm that increasing input variety minimises model overfitting and enables the model to generate consistent, applicable, domain-specific safety measures. The proposed methodology provides a highly scalable solution for automated risk management in high-risk industrial environments, such as mining, offering a reliable, data-driven decision-support mechanism that minimises the limitations of manual review.

Keywords: coal mining; data augmentation; large language models; low-resource fine-tuning; mining safety; Mistral-7B; occupational health and safety; precaution generation; QLoRA

Check for updates

Academic Editors: Nicola Magnavita and Guangjie Han

Received: 29 March 2026
Revised: 20 May 2026
Accepted: 4 June 2026
Published: 8 June 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

The mining industry is among the highest-risk sectors globally due to its inherent geotechnical, environmental, and operational hazards [1,2]. Risk assessments, accident investigation reports, and prevention texts are critical for protecting worker health and

reducing accidents in such hazardous environments [3,4]. Accident and root cause reports, meticulously documented by institutions such as the United States Mine Safety and Health Administration (MSHA), are invaluable sources of information that enable the prevention of future incidents and the development of proactive safety strategies [4]. Therefore, in-depth analysis of unstructured safety data is essential for building a safe working ecosystem [5].

Traditionally, the analysis of mine accident reports and risk texts has relied heavily on human expertise and manual review processes. However, manually analysing these texts, each of which can be tens of pages long, is an extremely labour-intensive and time-consuming process, far from practical [4]. Furthermore, rule-based approaches and traditional statistical text-mining methods are insufficient for capturing latent themes in complex occupational accident scenarios [6]. Traditional risk assessment methods can also lead to subjective judgments, expert biases, and operational delays (hysteresis) [2]. Faced with a multi-source, heterogeneous, and ever-increasing volume of data, the limits of manual and rule-based text analysis processes have been reached [2].

The dramatic advances in Large Language Model (LLM) and Natural Language Processing (NLP) technologies over the past few years offer revolutionary potential for extracting meaningful insights from unstructured safety reports [4]. Modern LLMs such as GPT-4 and LLaMA demonstrate powerful capabilities in rapidly processing hundreds of safety documents, categorising them while preserving semantic integrity, and presenting root causes in accident reports in interpretable, coherent summaries [4,7]. These models offer a unique opportunity to transform mine safety processes from reactive responses to data-driven decision -support systems that proactively predict hazards [2].

Despite the success of LLMs in general NLP tasks, their direct application to highly technical mining domains (e.g., “highwall,” “rib,” “face”) often results in semantic ambiguities and misclassifications [4]. Fine-tuning with domain-specific texts is essential for domain-specific adaptation of language models. However, mining safety texts fall into the category of “low-resource” datasets where high-quality, labelled data is limited [8]. Training LLMs with billions of parameters on limited data presents significant challenges, including high hardware costs and model overfitting. Even with modern techniques such as parameter-efficient fine-tuning (PEFT), the lack of annotations and data scarcity make it difficult to teach the model specific industrial terms [8].

A review of the existing literature reveals that the use of LLMs in mining safety is largely limited to passive text mining; for example, extracting hidden themes from MSHA reports [4] or classifying accident causes. However, the automated generation of an actionable, context-specific safety-precaution text (conditional text generation) in response to a defined risk scenario (e.g., “Production + Conveyor Maintenance + Crash Hazard”) remains a gap that has not yet been systematically addressed in the literature. Furthermore, the low-resource problem in this area is twofold: (i) the quantitative scarcity of labelled training data (n = 228 in this study) and (ii) the semantic deficiency or ambiguity of terms such as “rib”, “face”, and “highwall” in the general LLM vocabulary. This study targets precisely this gap, investigating how a model that not only classifies risk but also generates proactive precaution text by interpreting the operational context can be optimised under low-resource constraints.

Data augmentation strategies are strongly needed to improve the performance and generalisation ability of mine safety language models in low-resource settings [8]. When training LLMs with niche mining terminology, enriching limited datasets with synthetic data, zero or few-shot learning, or domain-specific dictionaries is a critical threshold for the model to accurately understand accident mechanisms and reduce false-positive/negative rates [8].

In light of the gaps identified above, the core research question of this study is sharpened as follows: "In low-resource mining security datasets, how can data-boosting-assisted parameter-efficient fine-tuning strategies be optimised to improve the performance of LLMs in generating context-sensitive safety measure texts, and what are the counter-effects of input- and output-side boosting techniques?"

This research adopts a different perspective from conventional LLM fine-tuning studies by focusing on the generation of conditional safety measures in a low-resource, operationally deterministic industrial environment. The proposed framework distinguishes itself from previous work in three main respects. First, it addresses a structured industrial production task where the set of valid outputs is inherently restricted by established safety protocols, as opposed to the open-field generative tasks usually investigated in the LLM literature. The second section focuses on how augmentation strategy design influences learning dynamics in low-resource settings, with a systematic contrast between input-side and output-side augmentation using identical QLoRA-based parameter-efficient fine-tuning. Third, the study provides empirical evidence that output-side multi-reference augmentation can compromise reliability in deterministic, safety-critical scenarios by introducing artificial uncertainty into the output distribution. To the best of our knowledge, this nuanced relationship between augmentation strategy and low-resource, safety-oriented conditional generation has not been systematically explored in previous industrial LLM adaptation studies.

The methodological originality of this study lies in a single, goal-oriented design choice: treating the choice of augmentation strategy not as a preprocessing detail, but as a fundamental research variable in itself within a simultaneously low-resource, domain-specific, and output-determining task. Unlike standard fine-tuning studies that apply augmentation uniformly or evaluate it solely in terms of data volume, this study systematically isolates the structural effect of the input and output spaces under the same model architecture, hyperparameters, and evaluation conditions. This design allows for a direct causal comparison rarely achieved in industrial NLP adaptation studies.

In this context, the study offers three significant contributions. First, it aims to create actionable conditional texts that go beyond classification and issue modeling by establishing the first end-to-end, QLoRA-based safety mitigation framework evaluated on a structured industrial risk dataset. Second, and most importantly, it empirically demonstrates that output-side multiple reference augmentation (E3), a strategy widely considered useful in the NLG literature, leads to a statistically significant performance degradation (BLEU: 16.02 $\rightarrow$ 12.21; BERTScore F1: 0.360 $\rightarrow$ 0.341) in low-resource, output-determining industrial environments. This negative finding is not coincidental: it stems directly from the low-entropy nature of safety-critical output domains and is corroborated by both quantitative metrics and qualitative output analysis. Third, it provides statistically validated evidence that input-side lexical enrichment (E2) delivers a 47% relative improvement (0.360 $\rightarrow$ 0.530) in BERTScore F1 because it increases input diversity while preserving output determinism; this is a design principle with direct implications for enrichment strategy selection in any domain where procedural correctness constrains the acceptable output space.

Related Work

In the fields of mining and occupational health and safety (OHS), systematic analysis of accident reports and compensation data is critical for identifying accident root causes and proactively mitigating risks [4]. These processes, traditionally requiring intensive human labour and time, are being automated in recent years through the integration of NLP and Machine Learning (ML) algorithms [4]. To autonomously extract latent themes

from unstructured accident/fatality reports in the United States Mine Safety and Health Administration (MSHA) databases, hybrid topic modelling techniques such as Latent Dirichlet Allocation (LDA) with TF-IDF weighting, and LLMs such as GPT-4o are being used [4]. Similarly, in studies using machine learning and statistical approaches to analyse worker compensation claims in the Alaska mining industry, regression and random forest algorithms have proven successful at classifying injury severity and associated risk factors [9]. Another study, based on National Institute for Occupational Safety and Health (NIOSH) data, leveraged the effectiveness of ensemble ML-based predictive algorithms in predicting fatal accidents in the mining industry [10]. Despite these automation processes, the lack of rich, domain-specific datasets and dictionaries for OSH (Occupational Safety and Health) remains an area for improvement in text analytics research [4,11].

One of the most important mechanisms for enhancing LLMs' zero-shot generalisation capabilities is instruction tuning [12,13]. However, integrating general-purpose trained models into specialised domains such as finance, law, health, and natural sciences requires incorporating domain-specific semantic features into the model [12,13]. Researchers have introduced the Scientific Instruction Generation (SIG) model, which autonomously generates instruction-response pairings from scientific texts, to tailor LLMs to specific domains, and have developed targeted models such as the "DARWIN" LLM series [14]. New metrics, such as the "Task-Semantic Alignment Score (TSAS)", have been developed to measure how well model outputs align with human intent and domain ethics in multi-domain dialogue generation [12]. In the context of industrial use cases, to ensure that systems similar to automotive repair assistants adhere to complex domain-specific safety rules, command-tuning techniques such as Revision with Extracted Rules (RER) are applied, and the model's compliance with the rules is optimised [15]. Multi-stage training frameworks based on Supervised Fine-Tuning (SFT) and Direct Preference Optimisation (DPO), used for domain adaptation, prevent catastrophic forgetting of general language skills while ensuring safe, domain-specific productivity in fields such as biobanking and public health [16]. Custom datasets designed with parameters such as cultural contexts, values, and security principles in mind also play an indispensable role in alignment processes and LLM adaptation [17].

In low-resource languages with limited labelled data or for specific tasks, data augmentation is critical for AI models to overcome data scarcity [18]. Classical augmentation techniques in Natural Language Processing (NLP), such as synonym replacement at the token level or paraphrasing or conditional generation at the sentence level, are being reshaped by powerful LLMs today [18]. Frameworks like "CoDa" are designed to ensure that augmented texts remain consistent with the original data distribution in low-resource text classification applications. These frameworks perform quality-oriented data generation by applying strict constraints at the lexical, syntactic, length, and conceptual levels [19]. On the other hand, methods that use LLM-based distance supervision and automated text generation (ATG-DS) mechanisms when generating data with low-resource tags, and then perform ranking-based selection to filter out noise (Self-RDGS), improve performance in sensitive areas such as Relation Extraction [20]. In the challenges of multi-label classification, synthetic data augmentation techniques supported by conditional generators (LD-VAE, etc.) that strengthen the ability to adapt to new combinations (compositional generalisation) are used [21]. Low-resource languages like Amharic and Swedish have been found to significantly improve traditional NLP metrics through the use of targeted word replacement (TSSR) and ChatGPT supported synthetic text generation techniques that incorporate semantic context and POS (word type) constraints, as evidenced by research [22].

2. Methods

2.1. Overall Methodological Framework

This study presents a holistic methodological framework for systematically investigating how LLMs can be optimised on a low-resource dataset to generate automated precautionary text from structured mine safety risk records. The proposed framework consists of five main sequential steps and is visualised in Figure 1:

1. **Data Preparation and Splitting:** 228 structured risk records from coal mine operations were cleaned and split into training (70%, $n = 158$), validation (15%, $n = 35$), and testing (15%, $n = 35$) sets to prevent potential data leakage.
2. **Application of Data Augmentation Strategies:** Data augmentation was applied to the training set only, yielding four different experimental configurations: (E1) no augmentation (original data), (E2) input-side lexical augmentation, (E3) output-side multiple-reference augmentation, and (E4) combined augmentation. The details of implementing these strategies are described in Section 2.5.
3. **Parameter-Efficient Fine-Tuning:** For each augmented training set, the Mistral-7B-Instruct-v0.2 model was fine-tuned using the QLoRA method. In this phase, the model's base weights were quantised to 4 bits, and only low-rank adapter matrices were trained.
4. **Repeated Training and Evaluation:** To control for randomness in the training process, model training was repeated with five different random seeds for each experimental configuration. The performance of each trained model was measured on a fixed test set using BLEU, ROUGE, METEOR, and BERTScore metrics.
5. **Statistical and Qualitative Analysis:** The significance of differences in performance between different data augmentation strategies was evaluated using paired *t*-tests and Wilcoxon signed-rank tests. Additionally, the measure texts generated by the model were subjected to qualitative analysis for contextual accuracy and applicability.

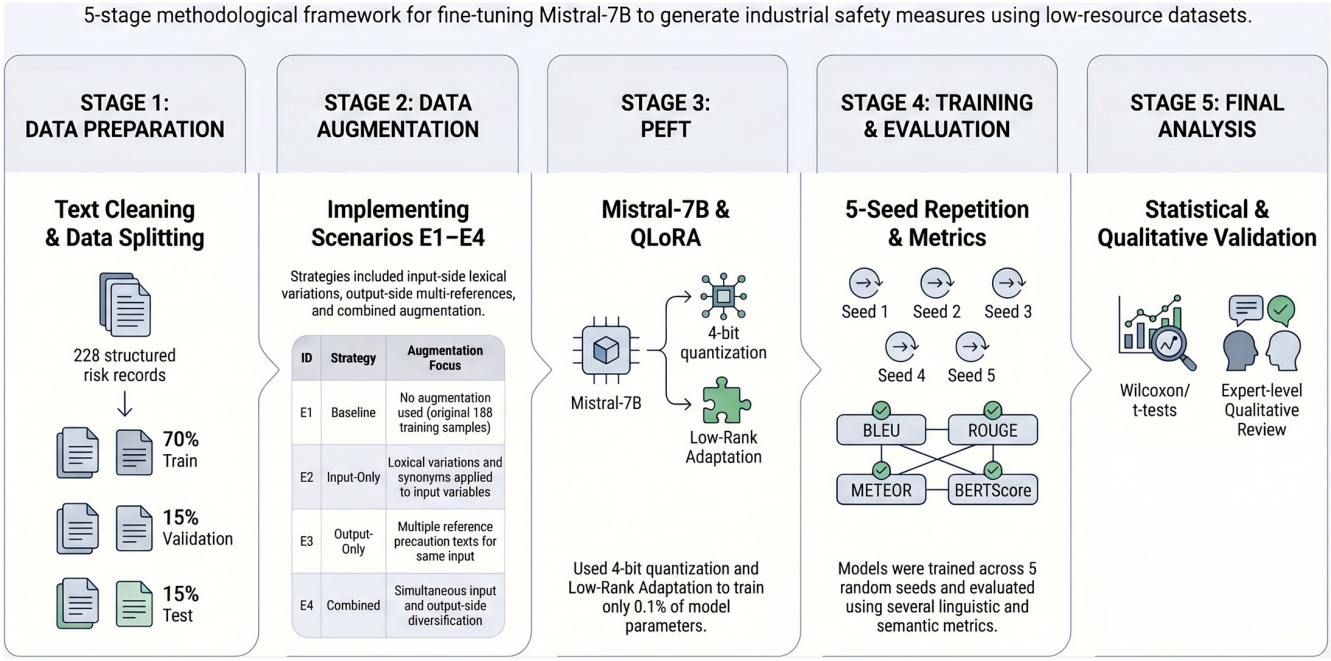


Figure 1. System Architecture.

The goal of this systematic framework is to compare the performance of diverse data augmentation strategies and provide a reproducible methodology for reliably deploying LLMs in a low-resource domain-specific NLP problem.

2.2. Problem Definition

The mining sector is considered one of the most critical industries globally in terms of occupational health and safety due to the high-risk working environments inherent in both underground and surface operations [23]. Various mechanical, environmental, and operational risks can arise during operations in underground and open-pit mining, leading to serious injuries, equipment damage, or production losses [24]. Therefore, systematically analysing risks and determining appropriate safety measures in mining activities is vital [25]. In risk assessment processes widely used in the industry, potential hazards and their consequences are generally recorded in structured records [26]. These records mostly consist of textual fields that define specific operational contexts, activities performed, equipment used, and potential risks. A crucial component of such records is recommendations for preventive measures to address the identified risks. However, manually generating these precautions is a process that relies on expert knowledge, is time-consuming, and prone to errors. Producing appropriate and consistent precaution texts for each risk scenario, especially in large-scale operations, presents a significant challenge [27]. Although the literature discussed in the previous section has presented various NLP and machine learning approaches to address this need, the development of a system that directly generates actionable security measures from structured risk records has not yet been systematically addressed.

In recent years, the success of LLMs in natural language generation has created significant opportunities to automate domain-specific text generation tasks [28]. LLM-based approaches can generate meaningful, consistent texts from a given context. Thanks to these features, it is possible to develop systems that automatically generate safety precautions based on risk definitions and operational context information [29]. The problem addressed in this study is to develop a language model that can generate appropriate precaution texts from structured risk records used in coal mining operations. In the dataset used in the study, each record consists of a set of independent variables that define the potential risks that may arise in a mining operation and the context in which they occur. The model must generate a logical, context-appropriate, and actionable precautionary text from these structured risk definitions. This is a conditional text generation problem, formally defined in Equation (1). The model's input space consists of the following fields that define an operational risk scenario:

- Unit: The mining unit where the operation is performed.
- Work_Done: The work or activity performed.
- Threat: The identified potential hazard.
- Equipment_Used: The equipment used during the operation.
- Risk: A description of the risk that the hazard may pose.
- Result: The possible outcome that may occur when the risk materialises.
- The expected output of the model represents the following area:
- Precaution: A safety measure that can be implemented to prevent or mitigate the effect of the identified risk.

Therefore, the model learns the following transformation. This transformation can be formally expressed as shown in Equation (1):

$$(\text{Unit}, \text{Work_Done}, \text{Threat}, \text{Equipment_Used}, \text{Risk}, \text{Result}) \rightarrow \text{Precaution} \quad (1)$$

In other words, the model learns to analyse a structured safety scenario, expressed through operational context and risk definitions, and generate a response text appropriate to this scenario. This task has significant practical applications in occupational safety and risk management. Automated response generation systems can support safety experts in

risk assessment processes, increase the consistency of recommended measures, and provide rapid recommendations for new risk scenarios. Especially in industrial environments where data-driven safety management approaches are becoming increasingly important, such systems can help develop decision-support mechanisms. However, domain-specific safety datasets are generally small, making it difficult to directly train LLMs. As a result, finding robust data augmentation methods to boost model effectiveness with limited, specialised datasets emerges as a key research challenge. In this work, we conduct a systematic investigation into how various augmentation strategies influence the generation of safety responses.

2.3. Dataset Description

The dataset used in this study was created from structured safety records obtained from risk assessment studies conducted in coal mining operations. The dataset consists of textual descriptions of potential hazards identified during field operations, the risks they may pose, and the recommended safety measures to address these risks. Such records constitute a crucial component of occupational health and safety management systems in the mining sector and enable the systematic monitoring and evaluation of operational risks. The dataset contains 228 records. Each record consists of six independent variables that define a specific operational context and one dependent variable that represents the recommended safety measure to be implemented within that context. The dataset structure is summarised in Table 1.

Table 1. Dataset structure.

Area	Description
Unit	The mining unit where the operation was carried out
Work Done	The work or activity performed
Threat	Identified potential hazard
Equipment Used	Equipment used during operations
Risk	Explanation of the risk that the hazard may pose.
Result	The consequences that may arise when the risk materializes.
Precaution	Recommended measure to prevent the risk or reduce its impact.

When these variables are considered together, each record represents a specific operational risk scenario. For example, a record might describe a potential hazard posed by equipment used during an activity in a particular mining unit, and the possible consequences of that hazard. In this context, the Precaution field represents the safety measure that should be implemented for that risk scenario.

The text fields in the dataset generally consist of short but information-rich statements. The texts in the input fields contain operational context and risk definitions, and mostly consist of a few words or short sentences. In contrast, the texts in the Precaution field generally contain more descriptive statements that define specific safety procedures or operational measures. This structure requires the model to understand the operational context and produce a precautionary text appropriate to it. Some data cleaning and editing operations were performed during the dataset creation process. These operations are summarised below:

- Incomplete or irrelevant records were removed from the dataset.
- Spelling inconsistencies and character encoding problems in the text fields were corrected.
- Field names and data formats were standardised.
- Duplicate records were checked and removed from the dataset.

As a result of these operations, the dataset was made consistent and usable for model training.

One important feature of the dataset is its domain-specific structure. Mining operation-specific terminology, equipment names, and operational processes are heavily present in the dataset. This situation can make it difficult for general-purpose language models to directly achieve high performance on this data. Therefore, appropriate fine-tuning and data augmentation strategies must be used to enable the model to learn domain-specific contexts.

Another important feature of the dataset is relatively small scale. The limited total number of records poses a significant challenge, especially for training LLMs. Small datasets increase the risk of overfitting the model and can limit its generalisation ability. Therefore, in this study, various data augmentation strategies were developed to make the dataset more effective in the training process. To more concretely illustrate the structure of the dataset, an example record is presented in Table 2.

Table 2. An example record from the dataset.

Unit	Work Done	Threat	Equipment Used	Risk	Result	Precaution
Production	Conveyor maintenance	Mechanical entrapment	Conveyor belt	Worker hand caught in moving parts	Serious injury	Lock out the conveyor system and ensure power isolation before maintenance operations

As shown in Table 2, each record in the dataset represents a specific operational risk scenario and includes the safety measure that should be implemented for that scenario. The model's task is to analyse these structured risk descriptions and generate an appropriate and context-sensitive Precaution text. This dataset has high practical value because it is based on real operational risk records. However, the limited size of the dataset and the inclusion of domain-specific terminology also present some challenges in model training. Therefore, in the following sections of the study, different data augmentation strategies aimed at improving model performance on small datasets are systematically examined.

2.4. Data Splitting and Leakage Prevention

To evaluate model performance reliably in machine learning-based text generation studies, it is crucial to split the dataset accurately and prevent potential data leakage during the training process [30]. Data leakage occurs when the model acquires direct or indirect information about the test data during training, potentially leading to an overestimation of model performance [31]. Therefore, the data splitting process in this study was carefully designed, and various control mechanisms were implemented to prevent data leakage. The dataset used in the study was divided into three subsets: training, validation, and test datasets. The goal of the data splitting process was to precisely separate the data used in the training process from the data used in performance evaluation. The dataset was split into test (15%) and validation (15%) sets using a fixed random seed (random_state = 42). Stratification based on the Unit variable was applied to preserve the distribution of potential class imbalance between splits. These parameters were fixed to ensure the reproducibility of the experiments. The distribution obtained after dataset splitting is shown in Table 3.

Table 3. Dataset distribution table.

Dataset	Number of Samples
Train	158
Validation	35
Test	35
Total	228

A significant methodological decision was made during the data splitting process: data augmentation operations were applied after the data splitting. In other words, the dataset was first split into training, validation, and test subsets, and then data augmentation operations were performed only on the training dataset. The main purpose of this approach is to prevent the samples created as a result of data augmentation from leaking into the validation or test datasets. To prevent data leakage, the following measures were implemented during the data splitting process:

- Pre-Augmentation Data Splitting: Before performing data augmentation, the dataset was fixedly split into training, validation, and test subsets. This ensures that the new samples created as a result of data augmentation remain only within the training data. This method is critically important in preventing indirect data leakage that may occur, especially in techniques such as paraphrasing or lexical augmentation.
- Context-Based Separation (Context Isolation): The records in the dataset represent specific operational contexts. Therefore, during the data splitting process, care was taken to ensure that different variations of the same context did not fall into different datasets. Thus, the model was prevented from indirectly acquiring information about the scenarios in the test dataset during training.
- Checking for Duplicate Records: Text-based similarity checks were applied to detect possible duplicates in the dataset. As a result of these checks, it was ensured that records with the same or highly similar content did not appear in different datasets. This process is important to prevent performance bias, especially in datasets consisting of short texts.
- Fixed Test Set: The same test dataset was used throughout the experiments. This approach ensures fair and comparable results among different data augmentation strategies. In all experiments, the model was trained only with training data, and hyperparameter adjustments were performed using the validation dataset.

Thanks to this data splitting strategy, any direct or indirect information transfer between the training, validation, and test datasets was prevented. Thus, it can be assumed that the obtained performance results reflect the model's true generalisation ability.

In conclusion, the data splitting and data leakage prevention strategy applied in this study is designed to be consistent with best practices recommended in NLP studies, especially those performed with small and domain-specific datasets. This approach provides a solid empirical foundation for reliably analysing the true impact of data augmentation techniques on model performance.

2.5. Data Augmentation Strategies

LLMs often require large and diverse datasets to be effectively trained on domain-specific tasks. However, the data collection process is often limited to domain-specific sources, such as industrial safety records, and the dataset remains relatively small. This situation increases the risk of overfitting, especially during fine-tuning of large-parameter language models, and can limit the model's generalisation performance. In such cases, data augmentation techniques are among the most commonly used approaches to improve

model performance. Data augmentation methods aim to increase the diversity of training data by creating new instances derived from the existing dataset [32].

In this study, four experimental setups were designed to analyse the effects of data augmentation strategies on the generation of safety measures using LLMs. In these experiments, data augmentation strategies were systematically applied, and the contribution of each strategy to model performance was compared. The experiments designed encompass the following four different data augmentation scenarios:

- Baseline model without data augmentation,
- Input-side lexical data augmentation,
- Output-side multi-reference data augmentation,
- Combined approach using both input and output-side data augmentation,

Through this approach, the effects of various data augmentation strategies on model performance in small, domain-specific datasets were systematically examined. In addition, in order to fairly evaluate the impact of the four data augmentation strategies described in the study on model performance, the following conditions were kept constant in all experiments:

- The same model architecture was used
- The same hyperparameters were applied
- The same training, validation, and test datasets were used
- Only the data augmentation strategy was changed

Based on this experimental design, it can be assumed that the observed performance differences are directly due to the data augmentation strategies.

2.5.1. Experiment 1—Training Without Data Augmentation (Baseline)

The first experiment is a baseline comparison experiment conducted without data augmentation. In this experiment, the model was trained only on the training samples in the original dataset. This approach aims to establish a baseline to evaluate the impact of data augmentation strategies on model performance. In this baseline experiment, the training dataset consisted solely of the original records, with no textual variation. The number of samples in the training dataset was 158.

2.5.2. Experiment 2—Input-Side Lexical Data Augmentation

In the second experiment, data augmentation was applied only to the input variables. In this approach, limited linguistic variation was introduced in the text fields defining the risk scenario. These variations were generally created using the following methods:

- Synonym replacement
- Minor paraphrase transformations
- Superficial language variations

These operations focused on text fields such as Work_Done, Threat, Risk, and Result. Since these fields define the operational context, it was assumed that the variations created in these fields would help the model learn the context more flexibly. A key feature of this approach is that the Precaution field was not modified. In other words, data augmentation was applied only on the input side, and each variation was matched with the same precautionary text. This design decision aims to enable the model to associate risk definitions with different wordings of the same safety precaution. The main goal of this approach is to improve generalisation performance by making the model more robust to linguistic variation. In this experiment, the training dataset contained 474 samples.

2.5.3. Experiment 3—Output-Side Multi-Reference Data Augmentation

In the third experiment, data augmentation was applied to the output text. This approach generated multiple alternative Precaution statements for the same input scenario. This method is based on the multi-reference augmentation approach frequently used in the natural language generation literature. In this strategy, different forms of precautionary texts were generated for the same operational risk scenario. For example, the same safety procedure can be rephrased using different sentence structures or word choices. This approach aims to enable the model to learn different forms of expression that carry the same meaning. As a result of this method, multiple output references were created for each input scenario, and the training dataset was expanded to 474 samples. This approach is widely used in the literature, particularly for natural language generation tasks, because it allows the model to produce a broader range of texts.

2.5.4. Experiment 4—Combined Data Augmentation Approach

The final experiment employed a combination of input-side and output-side data augmentation techniques, which involved creating lexical variations for the input texts and generating alternative precaution statements for the output texts. The combination of these two methods further enhanced the diversity of the training dataset. Thus, the aim was for the model to learn both different risk definition variations and different precaution statement formats. As a result of this data augmentation strategy, the training dataset reached 790 samples. This approach aims to maximise data diversity. However, using input and output variations together can sometimes complicate the model's learning process. Therefore, the effect of this strategy on model performance was evaluated experimentally.

2.5.5. Details Regarding the Implementation of Data Augmentation Strategies

To ensure reproducibility, this section describes the technical details and tools used in implementing data optimisation strategies. All optimisation processes were applied only to the training set after separating the training, validation, and testing phases. Optimisation processes were performed in the Python 3.10 environment using entirely deterministic rule-based and dictionary-based methods, without using any external LLM or API calls.

In the Input-Side Lexical Data Augmentation strategy (E2), only the argument texts (Unit, Work Done, Threat, Equipment Used, Risk, Result) were processed; the target variable Precaution was not modified. The optimisation process was performed in two phases:

- Domain-Specific Terminology Replacement: For mining-specific expressions, a predefined synonym/alternative dictionary (COLUMN_REPLACEMENTS) is used. This dictionary contains 2–3 alternatives for each variable corresponding to the original expression. For example:
    - Action Performed: "Gas measurement" → ["gas monitoring", "gas level measurement"]
    - Threat: "Absence of sensors" → ["sensor deficiency", "sensor unusable"]
    - Equipment Used: "Anemometer" → ["air velocity meter", "ventilation measuring device"]
- General Expression Modification: For general expressions that are not domain-specific, regular expression (regex) based modifications are applied using the light_phrase_variation() function. For example:
    - "should be" → ["must be", "should be", "needs to be"]
    - "to ensure" → ["to be sure", "to verify"]
    - "employees" → ["personnel", "employees"]

For each original training example, 2 augmented examples were created (input_aug_per_row = 2). By preserving the original examples in the training set (include_original_in_augmented_train = True), the total augmentation factor became 3x.

In the Output Side Multireference Data Augmentation strategy (E3), alternative versions of the Precaution text were created for each risk scenario while keeping the input variables constant. The build_multiref_precaution() function followed these steps:

- Paraphrasing: The original warning text has been broken down into clauses using periods (.), semicolons (;), and commas (,).
- Clause-Level Diversification: Paraphrasing changes have been applied to each clause using the paraphrase_precaution_clause() function. For example:
  - "To ensure" → ["To ensure", "To ensure is required", "The operation must ensure"]
  - "Stop the job" → ["Stop the job", "Stop the operation", "Suspend the task"]
- Order Change: In cases with 3 or more clauses, the order of the clauses has been changed without altering the meaning.

Two alternative mitigation texts were generated for each original training example (multiref_per_row = 2). By preserving the original examples, a total increment multiplier of 3× was obtained.

In the Combined Data Augmentation strategy (E4), the methods described in E2 and E3 were combined. Two variants were created on the input side, and two alternative measure texts were generated on the output side for each input variant. As a result of this cross-multiplication (2 × 2 = 4), the total increment multiplier, including the original samples, became 5× (Table 4).

Table 4. Data Augmentation Set Sizes and Augmentation Multipliers After Training.

Experiment	Strategy	Original Size	Augmented Size	Augmentation Multiplier
E1	No Augmentation	158	158	1×
E2	Input-Only Augmentation	158	474	3×
E3	Multiple Reference Outputs	158	474	3×
E4	Combined Augmentation	158	790	5×

To prevent data leakage, after the original dataset was split into training/validation/test sets, all augmentation operations were applied only to the training set. Thanks to the unique group_id values assigned to each instance, all variants derived from the same original scenario were kept in the same partition (training); and it was programmatically verified that there was no leakage to the validation and test sets (training ∩ validation = ∅, training ∩ test = ∅).

2.6. Model Architecture and Fine-Tuning Procedure

In this study, an open-source LLM was used to generate safety measure text from structured risk records. The Mistral-7B-Instruct architecture was chosen as the model. This model, with approximately seven billion parameters, has been widely used in NLP studies in recent years due to its high performance, especially in natural language understanding and generation tasks [33]. The Mistral architecture is a Transformer-based language model that offers higher computational efficiency compared to many models with similar parameter sizes, thanks to its advanced attention mechanisms and optimised training

strategies. In addition, the use of an instruction-tuned version of the model enables more successful results in conditional generation tasks such as text generation guided by specific task definitions [34,35]. However, direct retraining (full fine-tuning) of LLMs requires substantial computational resources. Therefore, in this study, the QLoRA approach, a parameter-efficient fine-tuning (PEFT) method, was used for model training. The QLoRA method was developed as a technique for training LLMs with low memory consumption. In this approach, model weights are stored in a low-bit representation, and learning is performed only through small adaptation layers rather than the entire model. Thus, the basic parameters of the model are kept constant, and learning is performed via LoRA adapters. This method enables training LLMs on domain-specific datasets, especially with limited hardware resources [36].

In this study, model weights were represented using 4-bit quantisation. In this way, GPU memory usage was significantly reduced, and the training process became more efficient. LoRA adapters were added to the model's attention layers, and the model learned task-specific information through them. The basic hyperparameters used in the model's fine-tuning process are shown in Table 5.

Table 5. Hyperparameters of the fine—tuned model.

Parameter	Value
Quantization Method	4-bit NF4
LoRA Rank	16
LoRA Alpha	32
LoRA Dropout	0.05
Learning Rate	2×10^{-4}
Number of Epochs	10
Batch Size	1
Gradient Accumulation	32
Maximum Sequence Length	768
Gradient Trimming	1.0
Optimisation Algorithm	Paged AdamW 8-bit

The hyperparameters presented in Table 5 were selected based on best practices in the literature and experimental pilot studies to prevent overfitting in a low-resource dataset and to ensure efficient training on limited hardware resources. All experiments were performed in Python 3.10 environment on an NVIDIA V100 16GB GPU using Transformers 4.41.2, PEFT 0.12.0, and TRL 0.9.6 libraries. The technical rationale for the selections is explained in detail below:

- **Quantisation Method (4-bit NF4):** The standard configuration of the QLoRA approach was used [36]. 4-bit NormalFloat (NF4) quantisation offers performance closest to 16-bit precision while preserving the distribution of model weights and reducing memory usage by approximately 4 times. Memory savings are further enhanced with double quantisation (double_quant = True).
- **LoRA Rank (r = 16) and Alpha (α = 32):** These values are commonly used in the QLoRA literature for 7B parameter models. Choosing a Rank value of 16 provides sufficient capacity for the model to learn domain-specific syntax and terminology, while minimising the risk of overlearning by limiting the number of trainable parameters (approximately 0.1% of the total parameters). Setting the α/r ratio to 2 is a standard approach to increase the learning signal of the adaptation layers.
- **LoRA Dropout (0.05):** A modest dropout was applied to the adaptation layers to prevent overtraining on low-resource datasets. This value is within the range recommended in the original QLoRA paper.

- **Target Modules:** All attention (query, key, value, output) and feedforward (gate, up, down) projection layers of the Mistral-7B model were targeted. This comprehensive selection maximises the model's domain-specific adaptation capacity.
- **Learning Rate (2×10^{-4}):** In fine-tuning studies with QLoRA, higher learning rates are tolerable compared to full model training. Pilot studies tested 1×10^{-4}, 2×10^{-4}, and 5×10^{-4}; 2×10^{-4} was observed to reduce validation loss most stably and quickly. The cosine learning rate program (lr_scheduler_type = "cosine") and a 5% warm-up rate (warmup_ratio = 0.05) prevented instabilities at the beginning of training.
- **Batch Size (1) and Gradient Accumulation (32):** This configuration is optimised for training a 4-bit quantized Mistral-7B model on an NVIDIA V100 GPU with 16 GB of VRAM. The low batch size addresses the memory constraint, while 32-step gradient accumulation effectively increases the batch size to 32, improving training stability and reducing gradient variance.
- **Number of Epochs (10):** It was observed that the validation loss plateaued after 10 periods. The model checkpoint with the lowest validation loss was saved for final evaluation using the 'load_best_model_at_end = True' parameter.
- **Maximum Sequence Length (768):** The token lengths of the combined input (prompt) and output (target) texts of all samples in the dataset were analysed using the Mistral-7B token. The 768 token value safely covers the longest sample in the dataset (95th percentile = 412 tokens) without causing any data loss (truncating).
- **Gradient Trimming (Maximum Gradient Norm = 1.0):** A standard value is used to improve training stability and prevent gradient bursts.
- **Optimisation Algorithm (Paged AdamW 8-bit):** A memory-efficient 8-bit optimizer compatible with QLoRA is used.

To control for variance arising from randomness, all experiments were repeated with 5 different random seed values (1, 2, 3, 4, 5). All code and configuration files will be publicly available as source code after the paper is accepted.

These hyperparameters were selected to ensure stable training on small datasets. Efficient training with limited GPU memory was achieved, particularly by using a low batch size and gradient accumulation strategy. During model training, each data record was presented to the model using a specific instruction format. In this format, input fields were combined into a structured contextual text, and the model was asked to generate a Precaution text appropriate to this context. Thus, the model learned to analyse input fields representing a risk scenario and generate the appropriate safety measure. During training, model outputs were regularly evaluated on the validation dataset, and model performance was monitored.

The fine-tuning process was followed by the evaluation of the model's performance on the test dataset, which was done using validation data to reduce the risk of overlearning and ensure stable progress during training. This setup ensured that outputs from various experimental conditions could be directly and fairly compared. The fine-tuning methodology adopted here supports effective adaptation of LLMs to compact, specialised datasets. By utilising the QLoRA-based training process, the model achieves a notable reduction in computational demands while maintaining its ability to acquire domain-relevant knowledge. This makes the approach especially well-suited for scalable and practical NLP applications that operate under data and resource constraints in industrial contexts.

2.7. Experimental Setup

To reliably and comparably analyse the impact of the proposed data augmentation strategies on the LLM's performance, the experimental process was systematically designed. The experimental setup uses the same model architecture, training parameters, and data

splitting strategy across all experiments to isolate the effect of different data augmentation strategies on model performance. Thus, it can be assumed that the observed performance differences are due solely to the data augmentation methods. The experimental evaluation process consists of three main components: repeated training with multiple seeds, evaluation metrics for measuring text generation performance, and statistical analysis of the experimental results.

2.7.1. Repeated Training with Multiple Seeds

In deep learning-based models, the training process involves a certain level of randomness due to randomly initialised parameters, data mixing operations, and optimisation processes. This can lead to different results even when training is performed using the same model and dataset. Therefore, it is recommended that experiments be repeated using multiple random initial values to more reliably evaluate model performance. In this study, model training for each data augmentation strategy was performed with five different random seeds. The seed values used are as follows:

$$\text{Seed} = \{1, 2, 3, 4, 5\}$$

For each seed, the model was trained from scratch, and the final outputs were evaluated on the test dataset. This approach prevents the model's performance from being dependent on a single training result and ensures more reliable results. When reporting the experimental results, mean and standard deviation values were calculated for each metric. In this way, the performance of different data augmentation strategies could be compared not only in terms of average success values but also in terms of performance stability.

2.7.2. Evaluation Metrics

In this study, BLEU, ROUGE, METEOR, and BERTScore, commonly used in the natural language generation literature, were used to assess the quality of the safety measure texts generated by the model. The main reason for using these metrics together is that evaluating model performance in text generation tasks with only a single metric is usually insufficient. Different metrics evaluate text similarity from different perspectives:

- BLEU (Bilingual Evaluation Understudy) metric measures the n-gram-based overlap between the text generated by the model and the reference text. This metric is particularly used to evaluate word order and superficial text similarity [37].
- ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metric measures the overlap between the reference text and the generated text in a recall-oriented way. It is a metric commonly used in summarising and text generation studies [38].
- METEOR (Metric for Evaluation of Translation with Explicit Ordering) metric evaluates word matches not only superficially but also by considering linguistic relationships such as synonymy and root similarity. This feature enables more flexible evaluation in text generation tasks [38].
- BERTScore, on the other hand, is a similarity metric based on contextual language models. This metric calculates the semantic similarity between the generated and reference texts using contextual word representations. Therefore, BERTScore is considered an effective measure, especially for evaluating semantic similarity [38].

Thanks to the combined use of these four metrics, model performance could be comprehensively evaluated in terms of both superficial word overlap and semantic similarity.

2.7.3. Statistical Analysis

Additional statistical analyses were performed to determine whether the performance differences observed among different data augmentation strategies were statistically significant. In machine learning experiments, comparing only average performance values is often insufficient, as observed differences may be due to random variation. Therefore, two different statistical tests were applied to compare the experimental results in this study:

- Paired *t*-test is a parametric test used to evaluate whether the difference between the average performances of two experimental groups is statistically significant. This test is commonly used when the data are approximately normally distributed.
- The Wilcoxon signed-rank test is used as a non-parametric alternative and can provide more reliable results when the data distribution is not normal. This test evaluates significance by analysing the median differences between paired measurements from two experimental groups [39,40].

Using these statistical tests, it was analysed whether the effect of different data augmentation strategies on model performance was not only observationally but also statistically significant.

2.7.4. Experimental Comparison Framework

To ensure fair and comparable testing, the following conditions were kept constant across all experiments: The same model architecture was used, the same fine-tuning method was applied, the same hyperparameters were used, the same training, validation, and test datasets were used.

The only factor changed between experiments was the data augmentation strategy. This allows us to assume that the performance differences obtained are directly due to the data augmentation methods. This experimental setup provides a robust methodological foundation for reliably analysing the impact of different data augmentation strategies on LLM performance in small and domain-specific datasets.

3. Results and Analysis

The effects of the proposed data augmentation strategies on the LLM's performance in generating safety measures are thoroughly analyzed in this section using both quantitative and qualitative analyses of the samples. First, the base model trained without data augmentation is presented to demonstrate the impact of dataset limitations on model performance. The model results obtained by applying data augmentation methods are then reported and compared with those of the base model. BLEU, ROUGE, METEOR, and BERTScore metrics were used in the performance evaluation. Furthermore, to increase the reliability of the experiments, each model was trained with five different random seeds, and the results were reported as mean performance and standard deviation. Additional statistical analyses were performed to determine whether the effect of data augmentation strategies on model performance was statistically significant. Finally, a qualitative analysis was conducted of the safety measure texts generated by the model to evaluate its capacity to produce contextually meaningful and applicable safety recommendations.

3.1. Baseline Model Performance

To reliably evaluate the impact of proposed data augmentation strategies on model performance, the baseline model trained without data augmentation was first analysed. In this experimental setup, the model was trained using only the original dataset and evaluated on the test dataset. This provided a reliable reference point for comparing the performance gains offered by data augmentation strategies. To reduce variation due to randomness, model training was repeated with five different random seeds. The model

outputs for each seed were evaluated on the test dataset, and performance was measured using BLEU, ROUGE, METEOR, and BERTScore. Final results are presented by reporting the mean and standard deviation for each metric.

As shown in Table 6, the model trained only on the original dataset showed limited text generation performance. In particular, the relatively low scores obtained in n-gram based metrics such as BLEU and ROUGE indicate that the predictive texts generated by the model did not achieve a high level of word overlap with the reference texts. The model's ability to learn a wider range of expressions was limited by the training dataset's limited size, indicating that it only achieved a moderate level of semantic agreement with the reference texts, according to the average F1 score in the BERTScore metric. However, the observed standard deviations in the metric values indicate some performance variation across training runs with different seed values. This indicates that model stability may be limited in language model fine-tuning performed with small datasets. According to these findings, LLMs trained on domain-specific, small-scale datasets may encounter difficulties in achieving sufficient generalization capacity in text generation tasks. Therefore, data augmentation strategies that enhance the diversity of the training dataset can improve text generation performance by aiding the model in learning a broader range of expressions.

Table 6. Baseline model performance (original dataset).

Metric	Mean	Std. Dev.
BLEU	16.02	2.15
ROUGE-L	0.377	0.056
METEOR	0.547	0.037
BERTScore (F1)	0.360	0.053

3.2. Impact of Data Augmentation

One of the most commonly used approaches in the literature to improve the performance of LLMs trained with small, domain-specific datasets is data augmentation techniques. In this study, an input-only data augmentation strategy based on paraphrasing and synonym transformations on input variables was applied to increase data diversity. In this approach, the independent variables in the dataset were rephrased to increase data diversity, while the target variable, the Precaution text, was preserved unchanged. Thus, the aim was for the model to learn to associate the same safety measure with different contextual expressions. To evaluate the effect of the data augmentation strategy on model performance, the model was retrained using the augmented dataset. To ensure the reliability of the experiments, the training process was repeated with five different random seeds, and the model outputs from each training session were evaluated on the test dataset. BLEU, ROUGE, METEOR, and BERTScore metrics were used in the performance evaluation. The final results are reported with mean performance and standard deviation values for each metric.

Table 7 shows that applying the data augmentation method led to a significant improvement in model performance. The notable increase in BLEU scores reflects better alignment of n-gram patterns between the generated and reference warning texts. Similarly, higher ROUGE-L and METEOR values indicate that the model better captures and replicates key structural elements found in the reference texts. The improvement in BERTScore, which assesses contextual similarity, shows that the data augmentation method helped the model grasp deeper semantic connections beyond superficial word matches.

Table 7. Model performance with input-only data augmentation.

Metric	Mean	Std. Dev.
BLEU	29.50	13.12
ROUGE-L	0.544	0.105
METEOR	0.683	0.065
BERTScore (F1)	0.530	0.113

This indicates that the model can produce more contextually consistent and meaningful safety measures. Overall, the data augmentation strategy has a significant impact on the model's text generation performance. Increasing data diversity, especially in small, domain-specific datasets, enables the model to learn diverse forms of expression, thereby significantly improving its generalisation capacity. These findings demonstrate that data augmentation methods are an effective approach to improving LLM performance in domain-specific text generation tasks.

3.3. Comparison of Augmentation Strategies

This section compares the effects of different data augmentation strategies on model performance. Four different experimental setups were evaluated in the experimental study: (i) the basic model trained without data augmentation (E1), (ii) a data augmentation strategy based on paraphrase and synonym transformations applied only to the input variables (E2), (iii) a multiple reference data augmentation approach based on generating multiple reference Precaution texts for the same input (E3), and (iv) a combined approach where both data augmentation strategies were applied together (E4). In all experiments, model training was repeated with five different random seeds, and the results were evaluated using BLEU, ROUGE-L, METEOR, and BERTScore metrics. Average performance values for each metric are presented in Table 8.

Table 8. Comparison of augmentation strategies.

Experiment	Augmentation Strategy	BLEU	ROUGE-L	METEOR	BERTScore (F1)
E1	No Augmentation	16.02	0.377	0.547	0.360
E2	Input-Only Augmentation	29.50	0.544	0.683	0.530
E3	Multiple Reference Outputs	12.21	0.356	0.521	0.341
E4	Combined Augmentation	14.33	0.372	0.538	0.355

Table 8 shows that data augmentation strategies have different effects on model performance. Specifically, the data augmentation approach (E2), applied only to input variables, provided the highest performance across all evaluation metrics. In this experiment, significant increases were observed in BLEU, ROUGE-L, METEOR, and BERTScore values compared to the base model. This result indicates that diversifying the input variables with different expressions helps the model learn contextual relationships more effectively. In contrast, the data augmentation strategy (E3), which generates multiple reference outputs, appears to decrease rather than improve model performance. This suggests that having multiple target texts for the same input can create uncertainty in the model's process. Especially in text generation tasks, high variation in the target variable can make it difficult for the model to learn the correct output distribution. The combined approach (E4), which applies both data augmentation strategies provided a small performance increase over the

base model but lagged behind the method that only enhanced input variables. This result shows that data augmentation strategies do not always complement each other and can, in some cases negatively impact model performance.

Overall, the findings indicate that data augmentation strategies that increase input diversity are more effective for text generation tasks using small, domain-specific datasets. Conversely, methods aimed at increasing target text diversity may introduce additional uncertainty into the model's learning process, thereby leading to performance degradation. These results show that designing data augmentation strategies, not only the increase in data volume but also its impacts on the model's learning process should be carefully considered.

To better understand the underlying mechanism of the performance degradation observed in the E3 strategy, the distributional characteristics of the target texts (Precaution) in the training sets have been conceptually examined. In the E1 and E2 strategies, there is only one reference measure text per input scenario; meaning the target output is deterministic within a given context. In contrast, the E3 strategy defines multiple target texts for the same input, each lexically and structurally different from the others. This leads to the conditional output distribution P(\text{Precaution} | \text{Context}) becoming artificially multimodal. As the model attempts to learn the multimodal target distribution, it encounters contradictory slopes, especially in the limited—data regime. This forces the model into a "decision" phase regarding its output, ultimately leading it to adopt more general expressions with low overlap with the reference texts. This mechanistic analysis is discussed in detail in Section 4.1.

3.4. Statistical Significance Analysis

To assess whether the differences in performance among various data augmentation strategies were meaningful, further statistical tests were conducted. In the context of machine learning, relying solely on average performance metrics can be misleading, since apparent differences might simply result from random parameter initialisation or fluctuations during training.

Therefore, statistical hypothesis tests were applied to assess the reliability of the experimental results. In this study, two statistical tests were used to analyse performance differences experimental setups: the paired *t*-test and the Wilcoxon signed-rank test. The paired *t*-test is a parametric test used to evaluate the significance of the mean difference between two paired samples, whereas the Wilcoxon signed-rank test, a nonparametric alternative, can provide more reliable results, especially with small sample sizes. The statistical significance threshold was set at $p < 0.05$ in all analyses. Table 9 presents the *p*-values for comparisons across different experimental setups for the four evaluation metrics.

The results show that different data augmentation strategies have statistically significant effects on model performance. In particular, the performance improvements observed in the METEOR and BERTScore metrics are largely statistically significant. This indicates that data augmentation strategies improve both the superficial and contextual similarity of the texts produced by the model. In the BLEU metric, the difference between experiments E1 and E2 is not statistically significant. However, the significant differences between experiments E2, E3, and E4 suggest that the data augmentation strategy applied to the input variables is more effective compared to other data augmentation methods. In the ROUGE-2 metric, the difference between E2 and E1 is quite close to the significance level ($p = 0.0517$). This suggests that the data augmentation strategy shows a strong tendency to increase text overlap, but the statistical power may be limited due to the small dataset. Overall, the statistical analysis reveals that the data augmentation strategy, particularly when applied to the input variables, significantly improves model performance. In contrast,

the data augmentation approach using multiple reference outputs was found to reduce model performance, and this reduction was statistically significant. These findings indicate that changes in the input and target variables during the design of data augmentation strategies can affect the model's learning process in different ways.

Table 9. Statistically significant analysis of augmentation strategies.

Metric	Comparison	Wilcoxon p	Paired t-Test p	Significance
BLEU	E1 vs. E2	0.0625	0.102	Not significant
	E1 vs. E3	0.0625	0.0010	Significant
	E1 vs. E4	0.0625	0.0090	Significant
	E2 vs. E3	0.0625	0.0266	Significant
	E2 vs. E4	0.0625	0.0285	Significant
ROUGE-2	E1 vs. E2	0.125	0.0517	Borderline
	E1 vs. E3	0.0625	0.0030	Significant
	E1 vs. E4	0.0625	0.0140	Significant
	E2 vs. E3	0.0625	0.0040	Significant
	E2 vs. E4	0.0625	0.0066	Significant
METEOR	E1 vs. E2	0.0625	0.0223	Significant
	E1 vs. E3	0.0625	0.00063	Significant
	E1 vs. E4	0.0625	0.0103	Significant
	E2 vs. E3	0.0625	0.00154	Significant
	E2 vs. E4	0.0625	0.00284	Significant
BERTScore	E1 vs. E2	0.125	0.049	Significant
	E1 vs. E3	0.0625	0.00063	Significant
	E1 vs. E4	0.0625	0.0055	Significant
	E2 vs. E3	0.0625	0.0039	Significant
	E2 vs. E4	0.0625	0.0045	Significant

3.5. Metric Distribution Analysis

The results obtained from the E2 experiment, the highest-performing scenario in the study, were statistically analysed to measure the model's predictive consistency, inter-metric relationships, and stability at different random seed states. This analysis forms the basis for verifying the reliability of the proposed model in a critical field such as mining. The overall distribution of metrics obtained across five seeds in the E2 experiment is presented in Figures 2 and 3. Box-plot analysis shows that the metrics cluster within a very narrow range. In particular, the consistent clustering of BERTScore F1 values around an average of 0.530 across the five seeds and the low variance demonstrate that the model's semantic inference success is not coincidental; rather, it reflects the structural stability of the learned representations.

The closeness of the median values for the ROUGE and METEOR metrics confirms that the model produces consistent predictive texts across both vocabulary and syntactic structure.

The relationships between the metrics were examined using the heat map in Figure 4 and the scatter plot in Figure 5. According to the analysis results, the strong positive linear correlation observed between ROUGE-1 and ROUGE-2 indicates that the model can generate not only individual words but also technical bigrams in a contextually appropriate manner. The high correlation between METEOR and BERTScore suggests that the model-generated measures exhibit deep semantic similarity to security protocols, extending beyond their dictionary meanings.

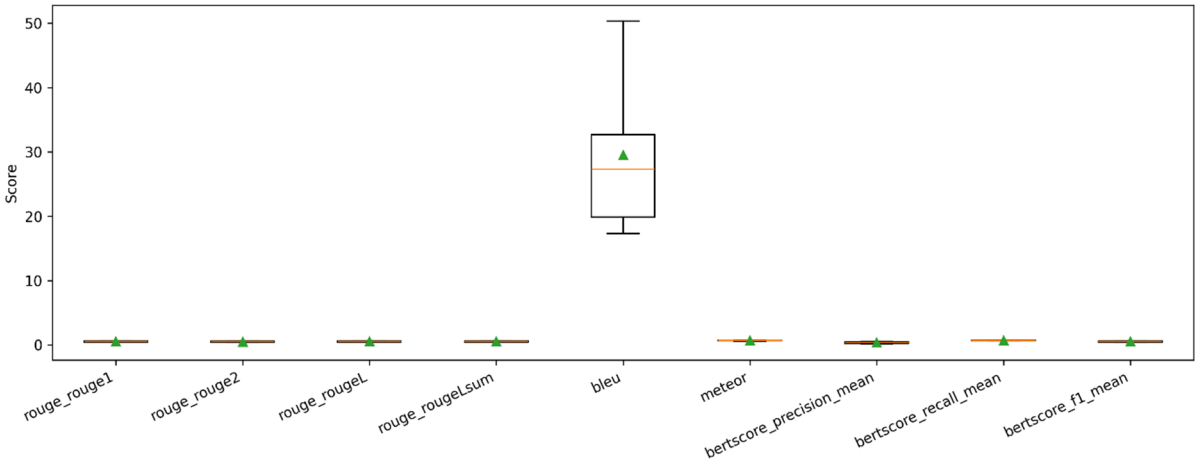


Figure 2. Metric distribution across seeds.

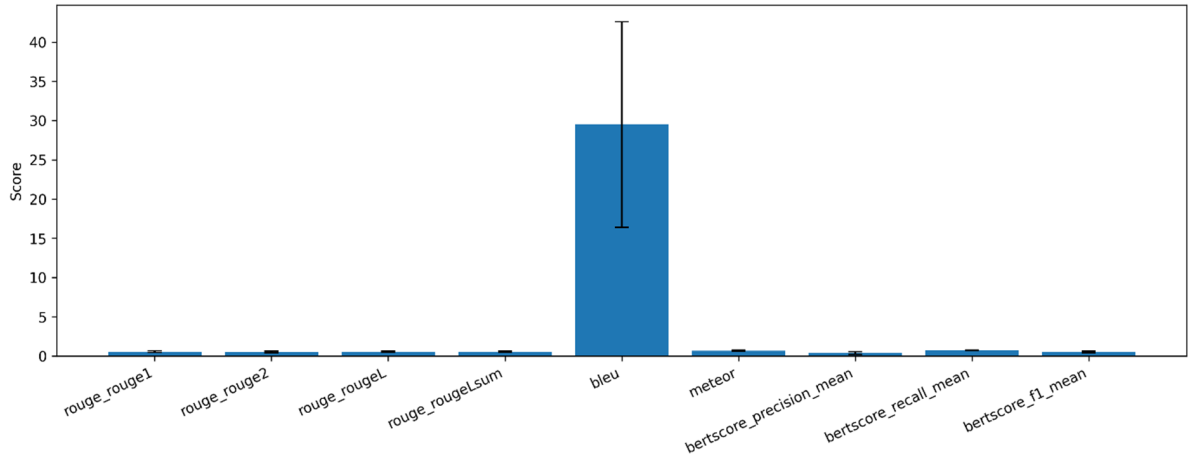


Figure 3. Mean ± std. dev. across seeds.

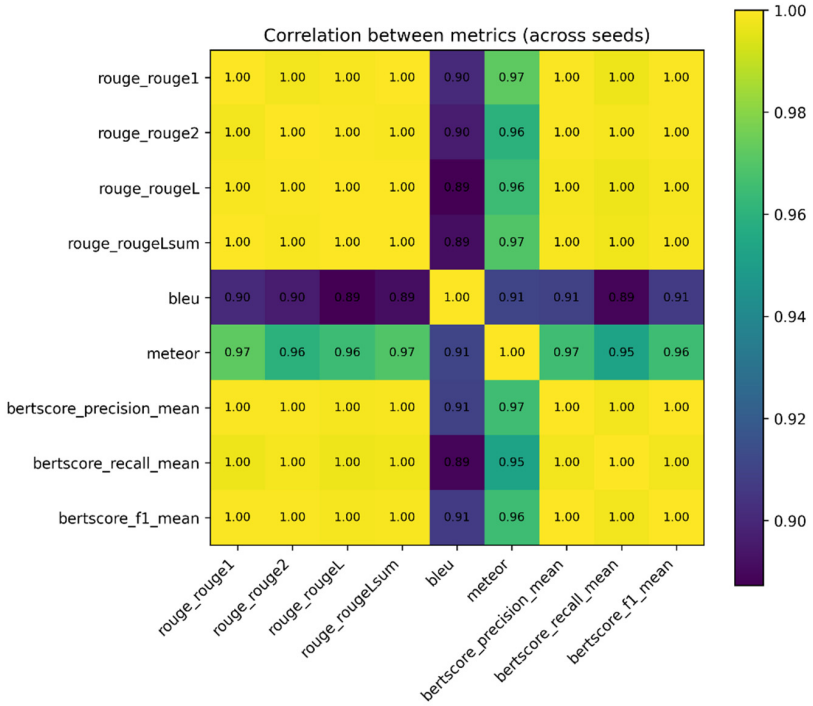


Figure 4. Correlation between metrics (across seeds).

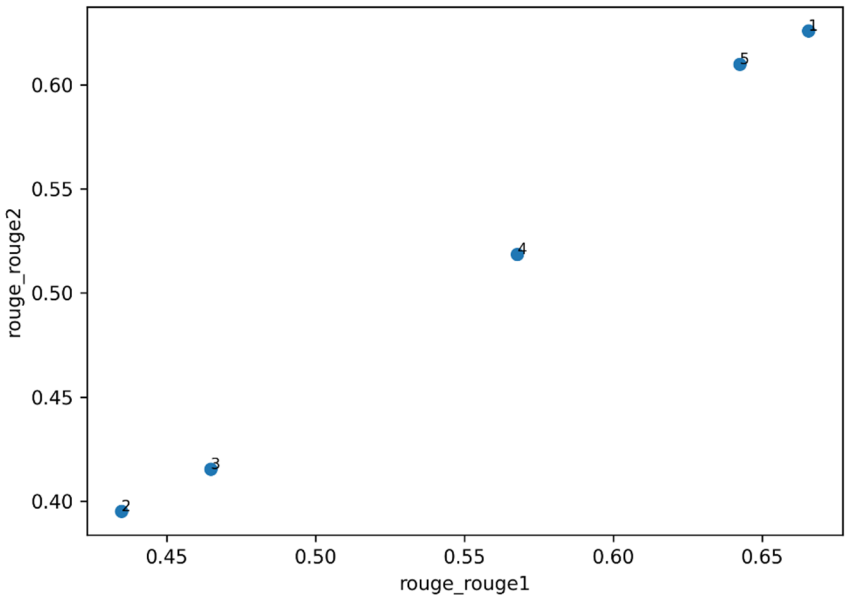


Figure 5. Each point is a seed (rouge2 vs. rouge1).

The histograms in Figure 6 reveal that the model performance exhibits a trend close to a normal distribution:

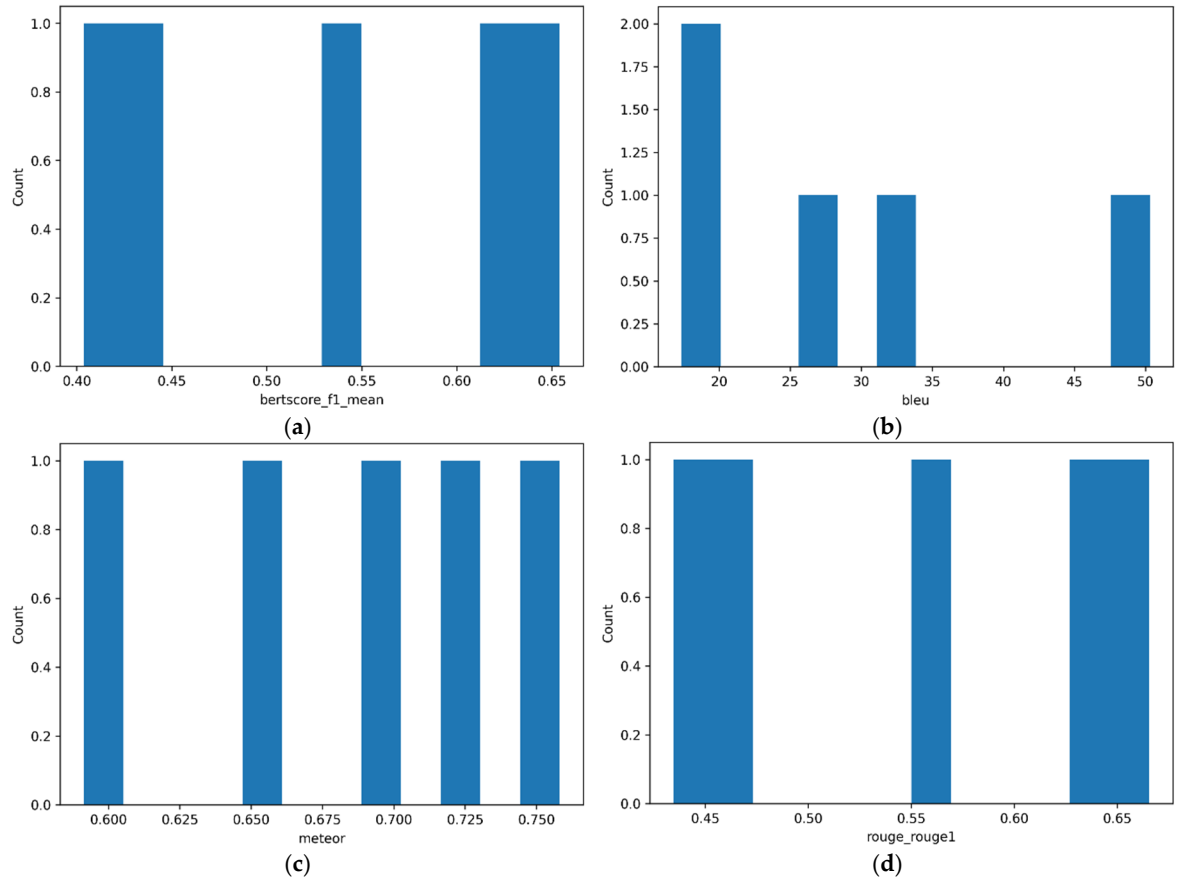


Figure 6. Metric distribution across seeds (**a**): bertscore f1 mean distribution; (**b**): bleu distribution; (**c**): meteor distribution; (**d**): rouge distribution.

BLEU Score: Concentration of the distribution at a specific frequency indicates that the model has successfully learned patterns in the reference texts.

BERT Score F1-Mean: The distribution exhibits a right-skewed structure with low dispersion and is concentrated around an average F1 value of 0.530. This distribution pattern reveals that low-quality predictions are kept to a minimum across the seeds, and that the semantic production quality of the model is stable and reproducible regardless of appropriate initial conditions.

METEOR Density: The high success rate of METEOR scores across a wide range, thanks to its flexible matching capability, supports the model's ability to use diverse but accurate terminology specific to different risk scenarios.

The statistical visualisations confirm that the data augmentation strategy (E2) applied to the input variables within the scope of the study optimised the model's learning capacity. Low standard deviation values indicate that the model is free of randomness and that the risk measures generated in coal mines can be reproduced with high accuracy in each iteration. This fully meets the high-performance predictable criterion, a critical requirement for integrating AI-based risk management systems into industrial environments.

3.6. Qualitative Analysis of Model Outputs

While numerical metrics give a broad measure of overall performance, qualitative analysis serves as a valuable complement by examining the relevance and real-world applicability of the model's generated responses in context. This section presents an exploratory, illustrative review by the authors to gain a preliminary understanding of the model's behaviour across different scenarios. It is important to note that this analysis does not include validation by independent mine safety experts and should therefore not be considered definitive proof of industrial applicability. In this review, model outputs from selected samples in the test set were categorised by the authors into three operational categories: (i) contextually correct and adequate (high semantic and technical overlap with the reference text), (ii) partially correct but incomplete or generalising, and (iii) erroneous or inadequate.

Contextually correct and sufficient outputs: The analysis results show that the model correctly understands the given risk context in many cases and generates appropriate and applicable safety measures. In particular, it was observed that the model produced outputs that closely overlapped with the reference texts in the risks associated with specific equipment use and operational processes. For example, in a dust control scenario, the model-generated precaution text correctly states that water -spraying systems should be used, as does the reference text. Similarly, regarding risks associated with sensor failures, the model suggests technically sound measures such as equipment inspection and replacement of faulty parts. Such examples demonstrate that the model can generate meaningful safety recommendations by understanding the operational context, rather than relying on superficial word matches. This finding is consistent with the high performance observed, particularly in the BERTScore metric.

Partially correct, generalised outputs: In some cases, the model captures the correct direction but does not achieve sufficient detail. Such outputs generally include general safety measures but do not include specific application details found in the reference texts. For example, in risks associated with cutting or excavation operations, the model generally produces statements such as "use of safety equipment" or "take appropriate protective measures." While these statements are technically correct, more specific applications (e.g., use of specific equipment or following specific procedures) found in the reference texts are not always included in the model outputs. This indicates that the model has learned general safety information but, in some cases, fails to adequately capture fine-grained operational details.

Erroneous and inadequate outputs: Although fewer in number, the model has been observed to produce measures that are not fully appropriate to the context or are incomplete.

Such errors usually occur in situations such as "input variables containing rare combinations," "scenarios not adequately represented in the dataset," and "complex structures where multiple risk factors are present simultaneously." In such cases, the model is seen to either use overly generalised expressions or omit some critical safety steps. This shows that the model's performance is directly related to the dataset's comprehensiveness.

Overall, the model is largely able to generate contextually meaningful and applicable safety measures. In particular, it has been observed that increasing the variety of inputs through data augmentation strategies enhances the model's ability to learn diverse expressions, thereby positively impacting the qualitative outputs. Nevertheless, in certain instances the model generates broader statements, and its effectiveness declines when confronted with less common cases. The results underscore that expanding data diversity enhances both quantitative performance measures and the contextual quality of outputs. Ultimately, this methodology offers a promising pathway toward building models capable of delivering relevant and actionable results in specialised text generation applications.

The examples presented in Table 10 were selected to examine the model's behaviour in different scenarios more closely. The results show that the model can generate technically accurate measures that highly overlap with the reference texts in many cases. In particular, in well-defined scenarios such as blasting operations, sensor failures, and excavation safety, the model outputs almost perfectly match the reference texts. However, in some cases, the model tends to use more general expressions. In scenarios that require more specific technical details, such as dust control and explosion prevention, the model accurately identifies the problem but falls short of the level of detail found in reference texts. These instances have been rated as "partially correct." Overall, these examples demonstrate that the model is largely capable of generating contextually meaningful and implementable safety measures, but in some cases requires a wider variety of data to learn more specific technical details.

Table 10. Example generated precautions and qualitative evaluation.

Scenario	Reference Precaution	Generated Precaution	Evaluation
Blasting operations & safety control	A guard is sent to each intersection point...	After the blast holes are connected, the guard is assigned to control the intersections...	Correct
Dust exposure (respirable dust)	For respirable dust control, a water sprinkler system...	Working in accordance with work instructions... dust control measures must be applied...	Partial
Face stability during excavation	A face board will be held against the face...	A face board will be held against the face before operation...	Correct
Sensor malfunction	Repairing faulty sensors, replacing them...	Repairing faulty sensors, replacing them...	Correct
Explosion risk (stone dust application)	Spreading impregnated stone dust in the shaft...	Spreading impregnated stone dust in the longwall...	Partial
Unsafe working conditions	No work orders can be given for unsafe points...	No work orders can be given for points where the risk continues...	Correct

4. Discussion

This study is one of the first comprehensive investigations to systematically examine parameter-efficient fine-tuning of LLMs for automated safety measure generation from structured risk records in high-risk industries such as coal mining, and the impact of data augmentation strategies in this process. The findings offer significant theoretical and practical implications for industrial NLP applications working with domain-specific and low-resource datasets.

4.1. Mechanistic Analysis of Output-Side Augmentation Failure

A key finding of this study is that output-side multi-reference augmentation (E3) statistically significantly reduced model performance (BLEU: from 16.02 to 12.21, BERTScore F1: from 0.360 to 0.341). This finding challenges the widely held assumption that more data invariably improves performance, particularly in low-resource, deterministic-output domains. In this section, possible mechanisms explaining the observed performance degradation are discussed within the framework of the task's specific characteristics and model learning dynamics.

4.1.1. Nature of the Task: Low-Entropy Output Space

In mining safety, risk assessment and mitigation rely heavily on standardised operational procedures and legal regulations. The structure of our dataset reflects this: the original training set (E1) contains only a single reference mitigation text for each unique input scenario (group_id). From an NLG perspective, this indicates that the conditional output distribution P(y | x) exhibits low entropy and a unimodal structure. In other words, given context x, the space for the "correct" values that output y can take is extremely narrow. This feature fundamentally distinguishes the task from tasks such as open-ended text generation or machine translation, where there are multiple valid outputs.

4.1.2. Artificial Entropy and Gradient Conflict Created by Multiple Referencing

The E3 strategy artificially transforms the naturally low-entropy output space into a high-entropy one by presenting multiple different y targets for the same input x. When autoregressive language models are trained with standard cross-entropy loss (\mathcal{L}_{CE}), they attempt to learn the empirical distribution \hat{P}(y | x) in the training data.

$$L_{CE} = \frac{-1}{N}\sum_{N}^{i=1} \log P_{\theta}(y_i|x_i) \quad (2)$$

In the E3 strategy, for a given input x_i, multiple targets y_i$^{(1)}$, y_i$^{(2)}$, . . ., y_i$^{(K)}$ are provided. The model is thus forced to learn a multimodal target distribution Pˆ(y | x_i), which can lead to the gradient conflict and probability mass dilution mechanisms discussed below.

This situation may lead to two mechanisms that degrade model performance:

1. **Gradient Conflict Hypothesis:** Different y_i$^{(K)}$ targets for the same x_i can generate gradients that pull the model parameters (\theta) in different directions. When these gradients are averaged, the net update signal may become weak or inconsistent. This phenomenon is similar to "negative transfer" in the multitasking learning literature. Although gradient analysis was not performed in this study, the performance degradation in E3 is consistent with this hypothesis.
2. **Dilution of Probability Mass Hypothesis:** When a model with limited capacity (especially a low-ranking adaptation like LoRA) is forced to distribute its probability mass across multiple different token sequences, the probability assigned to each sequence may decrease. This can make it difficult for the model to focus on the single most likely output when using greedy decoding or beam search during inference.

4.1.3. Effects Observed in the Context of Low-Resource Data

The impact of this potential learning conflict may have become more pronounced in the low-resource scenario, where training data is extremely limited ($N = 158$). The qualitative analysis presented in Section 3.6 revealed two types of undesirable behaviour in the outputs of the model trained in E3:

- Over-generalisation: The model tended to gravitate towards the lowest common denominator to reconcile the variability between different references, producing general

statements with low informational content, such as "take necessary safety precautions" (in the "partially correct" category).

- Inconsistency: In some cases, the model produced semantically inconsistent hybrid texts by combining parts from different sources.

These observations provide qualitative evidence that the increase in multiple references negatively affects the model's ability to produce consistent, specific outputs.

4.1.4. Comparative Perspective with the Success of E2

The success of the E2 strategy offers a complementary insight within this framework. E2 enriched the linguistic diversity to which the model was exposed by increasing the entropy of the input space, but preserved the deterministic structure of the output space ($P(y | x)$). This approach taught the model a many-to-one mapping, which can be summarised as "produce the same output for different inputs that mean the same thing". Empirical results show that this type of mapping is quite effective at increasing the model's generalisation ability and resistance to input variations in low-resource scenarios. This finding empirically confirms that input diversification is a safer and more effective strategy than output diversification, especially in low-resource areas.

4.1.5. Literature Context and Generalizability

This finding aligns with a growing awareness in the NLG literature: data augmentation strategies are sensitive to the nature of the task and the data regime. In machine translation, it has been reported that using one-to-one reference translations does not always improve the BLEU score, and performance can suffer when the references differ significantly. Similarly, in tasks such as image captioning, it has been shown that multiple references can reduce the model's originality. Our study provides new and robust empirical evidence to this debate from a low-resource, domain-specific, and security-critical context. Our findings suggest that data augmentation on the output side should be carefully considered, particularly in all fields where high accuracy and consistency of output are required, such as legal text analysis, medical reporting, and technical documentation.

4.2. The Impact of Data Augmentation Strategies on Performance

The quantitative results presented in Table 8 reveal a consistent and distinct pattern across all four evaluation metrics. The input-side lexical augmentation strategy (E2) outperformed all other configurations, increasing the BLEU score from 16.02 to 29.50 and the BERTScore F1 from 0.360 to 0.530 compared to the base model (E1). In contrast, the output-side multi-reference augmentation (E3) underperformed the base model in every metric (BLEU: 12.21; BERTScore F1: 0.341); and the combined strategy (E4) provided only limited improvement (BLEU: 14.33; BERTScore F1: 0.355). These differences were found to be statistically significant using both the paired *t*-test and the Wilcoxon signed-rank test, as shown in Table 9. The only exception is the E1-E2 comparison in the BLEU metric ($p = 0.102$). In this exception, insufficient statistical power (5 seeds, small sample) is thought to be the explanatory factor.

These empirical patterns are not merely dataset-specific observations; they reflect a deeper structural feature of the task. The conditional output distribution $P(y | x)$ in mine safety measure production is inherently low-entropy: acceptable outputs y for a given operational risk context x are strictly constrained by standardized safety protocols and legal requirements. Strategy E2 respects this structure by enriching the input distribution while not altering the output distribution; thus, it enables the model to perform many-to-one mapping learning that is both learnable and generalizable under conditions of data scarcity. Strategy E3, on the other hand, artificially imposed a multi-peak distribution on a struc-

turally monopeaked output space; this led to the gradient conflict and probability mass dilution mechanisms discussed in detail in Section 4.1. Scientifically, this finding makes a unique contribution to the data augmentation literature by demonstrating that augmentation effectiveness depends on the compatibility between augmentation design and the entropy structure of the target output space; a principle that has not been systematically addressed in previous low-resource industrial NLP studies. The failure of the combined approach (E4) reinforces this result: the addition of output-side augmentation consistently nullified the gains of input-side augmentation under conditions where the output space is deterministic; clearly demonstrating that the two strategies are not complementary.

4.3. Domain-Specific Performance of the Model and Practical Implications

The qualitative analysis presented in Section 3.6 reveals that the model trained with E2 exhibits high semantic accuracy in well-defined operational scenarios. In cases such as blasting operations, sensor failure, and excavation safety, the measures generated by the model closely matched the reference texts in both content and structure, with site-specific procedures and equipment requirements being correctly identified. Partial matches observed in scenarios such as dust control and explosion prevention indicate that the model correctly captured the safety category but failed to consistently reproduce the detailed procedural information in the reference texts. It is noteworthy that these partial outputs are concentrated in scenarios underrepresented in the training set; this suggests that the performance at this level stems primarily from the scope of the training set, rather than a fundamental limitation of the model architecture.

The above empirical findings have direct practical implications for the use of LLM-based tools in industrial safety management. First, the proposed framework should be best positioned as a decision support mechanism; it is designed to support, not replace, occupational safety professionals. In coal mining, where routine and well-defined operational scenarios constitute the vast majority of repetitive risk assessment tasks, the model reliably generates preliminary response drafts, thereby reducing documentation workload and increasing terminological consistency among operational units. Secondly, the scalability of the framework offers a significant practical advantage: the QLoRA-based training pipeline requires only a single NVIDIA V100 GPU with 16 GB of VRAM, making it accessible to industrial organizations without large-scale computing infrastructure. Thirdly, the input augmentation strategy (E2) is fully reproducible using deterministic rule-based methods without relying on external APIs or proprietary tools; this feature is critical for industrial deployment environments where auditability and reproducibility are regulatory requirements. Beyond coal mining, these practical features suggest that the proposed framework can serve as an adaptable template for other safety-critical sectors—construction, oil and gas, chemical processing—where structured operational risk records exist but domain-specific labeled datasets are limited.

4.4. Methodological Contribution and Orientation for Future Studies

The primary methodological contribution of this study goes beyond the application context and reveals a principle regarding the design logic of data augmentation in conditional text generation tasks. The model architecture, hyperparameters, and training/validation/test splits were kept constant, with only the augmentation strategy modified; thus, a controlled causality comparison was established that isolated the impact of augmentation design on learning dynamics. This empirical rigor, combined with iterative training with five different random seeds and paired statistical testing (paired t-test and Wilcoxon signed-rank test) [2,41], demonstrates that the observed performance differences are structural, not random. The resulting methodological principle—that augmentation

effectiveness is determined by the fit between the entropy of the augmented space and the entropy of the task's target output distribution—establishes a generalizable principle for augmentation strategy selection in low-resource industrial NLP. The reliability of this finding is further supported by the low standard deviation values obtained under the E2 strategy and the concentration of the BERTScore distribution at high performance values; Both indicators confirm that the model's gains are reproducible across training studies and are not a product of suitable initial conditions. These features meet the predictability and reproducibility criteria, which are prerequisites for the responsible integration of AI-based systems into safety-critical industrial environments [28].

In future studies, testing this methodology on datasets from different types of mining (open-pit, underground) or other high-risk sectors, such as construction, oil, and gas will increase the generalizability of the findings [27]. Furthermore, it is important to apply semantically richer data augmentation techniques (e.g., controlled paraphrase generation via LLMs) and to investigate their impact on the model's capacity to learn fine-grained details [22]. Finally, human-centred assessment studies involving occupational health and safety (OHS) experts to evaluate the practical applicability and compliance with industry standards of the safety measures developed will be an indispensable step toward the real-world integration of such systems [7,29].

5. Limitations and Future Work

Despite its methodological rigor, the findings of this study should be evaluated within certain limitations. These limitations are addressed below under four dimensions: scalability, generalizability, industrial deployment, and human-loop validation.

Scalability: The proposed framework was developed and evaluated on a 228-record dataset from a single underground coal mine, using a fixed model size (Mistral-7B) and a single GPU (NVIDIA V100, 16 GB VRAM). While designed to reduce QLoRA-based training memory requirements, the scalability of the framework has not yet been tested along three different axes. First, data scalability: it remains unclear whether the performance gains provided by the E2 strategy will be maintained, increased, or decreased when the training set reaches hundreds or thousands of records; as larger datasets may reduce the relative advantage of boosting by providing sufficient input variation through natural variation. Second, model scalability: the interaction between augmentation strategy design and model capacity has not been investigated; larger models (e.g., variants with 13B or 70B parameters) may respond differently to input-side augmentation due to their higher intrinsic generalization capabilities. Third, operational scalability: the current framework is a batch processing system evaluated in an offline environment. Scaling into a multi-user, real-time industrial environment where multiple users will request simultaneous response generation will bring with it inference latency, concurrent request management, and system reliability requirements; these requirements have not been addressed in the present study.

Generalisation Capability: The dataset used in this study consists of single-source risk records written exclusively in English and covering the operations of a specific underground coal mine. This single-source, monolingual design limits the generalizability of the findings in three ways. First, the operational terminology, risk scenario distribution, and mitigation formatting rules of this dataset may differ significantly from data obtained from different mine types (e.g., open-pit, metal, or salt mines), different sectors (e.g., construction, petrochemicals), or different regulatory jurisdictions. The effectiveness of the E2 strategy relies on the assumption that input variation is linguistically meaningful within a stable output space; this assumption may be invalidated where the safety protocols of the target area are less standardized or the acceptable output space is structurally broader. Second, the model was trained and evaluated only on English records. Regions with intensive mining

activities, such as Latin America, Eastern Europe, and East Asia, predominantly operate with languages other than English, and the morphological complexity of these languages may alter the effectiveness of rule-based lexical augmentation applied in E2. Third, the negative finding regarding E3—that output-side multi-reference augmentation degrades performance—was obtained in a single low-resource environment. While the mechanistic explanation presented in Section 4.1 suggests that this finding could be generalized to other deterministic output-space domains, this result cannot be accepted as a universal principle without empirical validation across different task structures and data regimes.

Industrial Deployment: The current study evaluates the proposed framework only in a controlled, offline experimental environment. Transitioning from this environment to active industrial deployment is outside the scope of the current study and presents a unique set of challenges that should be considered significant limitations. At the system integration level, deploying the model into existing safety management information systems (SMIS) or enterprise resource planning (ERP) platforms used in mining operations requires API development, pipeline engineering, and compliance testing. At the operational performance level, real-time response generation—possibly on the order of seconds per query in operational contexts—imposes response latency requirements necessitating model quantization, caching strategies, or hardware upgrades beyond the current V100 configuration. At the regulatory compliance level, AI-generated safety recommendations in high-risk industries are subject to varying occupational health and safety regulations by jurisdiction; many regulatory frameworks require automated outputs influencing safety decisions to be monitored, audited, and approved by certified safety professionals before operational use. Finally, at the user acceptance level, the practical utility of the system depends not only on output quality, but also on the extent to which occupational safety specialists trust, understand, and effectively interact with AI-generated recommendations—a dimension that requires dedicated human factors research and is not fully addressed in the current evaluation.

Human-in-the-Loop Validation: The qualitative analysis presented in Section 3.6 is an exploratory review conducted by the authors and does not constitute independent expert validation. This represents a significant limitation for a system intended for use in safety-critical industrial environments, as the operational adequacy of the generated measures cannot be validated solely by automated assessment metrics. High scores on BLEU, ROUGE, METEOR, and BERTScore confirm lexical and semantic similarity with reference texts; however, these metrics do not validate regulatory compliance, operational feasibility, or the absence of safety-critical deficiencies, which can only be assessed by certified occupational health and safety (OHS) professionals. Future validation studies should follow a structured, multi-stage human-loop protocol. This protocol should specifically include: (i) blind expert assessment, where independent OHS professionals evaluate the model outputs in terms of technical accuracy, completeness, and regulatory compliance without access to reference texts; (ii) cross-value reliability analysis to measure assessor agreement and identify patterns of systematic disagreement. (iii) scenario-based operational testing in which safety experts use the system in simulated risk assessment workflows and evaluate its practical utility, response appropriateness, and failure modes; and (iv) longitudinal auditing in which model outputs used in real operational contexts are retrospectively compared with event logs and the operational effectiveness of the measures suggested by the AI is assessed. Until such a validation protocol is completed, the system should be considered a decision support tool requiring mandatory expert review and approval at every stage of operational use. The present work does not include safety compliance strategies such as Human Feedback Reinforcement Learning (RLHF), Direct Reference Optimisation (DPO), constitutional AI, or output filtering mechanisms; the integration

and rigorous evaluation of these compliance techniques are a prerequisite for responsible operational deployment in high-risk industrial environments.

These limitations also offer concrete and valuable directions for future research:

1. **Cross-Industry Transfer Learning:** The performance of the model trained with the E2 strategy optimised in this study should be tested on new datasets collected from diverse but conceptually related fields such as open-pit mining, tunnel construction, or chemical plant maintenance. Specifically, the extent to which the model can adapt to these fields with a small number of new samples (few-shot learning) should be investigated.
2. **Multilingual Generalisation:** Testing the methodology on safety reports in languages of mining-intensive regions, such as Spanish, Russian, or Chinese, is critical to evaluating the approach's language independence and global applicability.
3. **Richer Data Augmentation Techniques:** Future studies could go beyond the rule-based lexical augmentation used in this research and explore domain-specific, finely tuned, smaller models or synthetic data generation techniques guided by human expert feedback. Finally, the rule-based lexical augmentation method employed in this study, while fully reproducible and computationally efficient, represents a relatively simple approach compared to advanced techniques such as LLM-based controlled paraphrase generation. Future work could systematically investigate the trade-offs between the simplicity and reproducibility of rule-based methods and the semantic richness offered by LLM-based augmentation in low-resource industrial NLP tasks.
4. **Safety Compliance and Rollback-Assisted Architectures:** Future research should also explore the integration of rollback-assisted generation (RAG), reinforcement learning from human feedback (RLHF), direct preference optimisation (DPO), and constitutional compliance strategies to enhance true consistency and reduce the risk of generating unsafe or operationally incompatible measures. Such compliance-oriented architectures may be particularly important for safety-critical industrial NLP applications where output reliability is more important than productivity diversity.

Aware of these limitations, we believe that the present study establishes a solid foundation in the field of low-resource industrial NLP and serves as a valuable reference point for future research.

6. Conclusions

This study makes two key academic contributions and offers one practical implication for deploying LLMs reliably and effectively in high-risk, low-resource domains such as coal mining.

Academic Contributions: The key methodological contribution of this study is the empirical demonstration that the effectiveness of the augmentation strategy is determined not only by the volume of data but also by the entropy structure of the target output space. In tasks where the conditional output distribution $P(y \mid x)$ is inherently low-entropy—such as in security-critical procedural text generation—artificially inflating the output entropy through multiple reference boosting (E3) leads to gradient conflict and probability mass dilution, resulting in statistically significant performance degradation across all four evaluation metrics. Conversely, maintaining output determinism while expanding input diversity (E2) yields consistent and statistically significant gains (a relative improvement of 47% in BERTScore F1: $0.360 \rightarrow 0.530$). This entropy-focused data augmentation principle is the study's primary contribution to the data augmentation literature and carries direct design implications beyond mining: wherever procedural compliance, legal certainty, or operational determinism constrains the acceptable output space (including medical docu-

mentation, legal text generation, and technical reporting), output-side data augmentation should be approached with equal care.

Practical Application Value: This study provides a viable roadmap for industrial organisations with limited amounts of domain-specific text data. It demonstrates that combining QLoRA-based parameter-efficient fine-tuning with a carefully designed input augmentation strategy can pave the way for a decision-support system that generates consistent, context-aware safeguards, potentially reducing reliance on costly human expertise. This approach is not limited to mining but serves as an adaptable template for other high-risk sectors (construction, energy, petrochemicals) with similar data constraints.

Future Studies: The findings of this study also open concrete pathways for future research. As detailed in Section 5, critical next steps include testing the generalizability of the proposed methodology to different industries and languages and subjecting the model outputs to a structured evaluation by independent domain experts. While acknowledging current limitations, we believe this study establishes a solid foundation in field of the low-resource industrial NLP and provides a valuable reference point for future research.

Author Contributions: Conceptualization, H.E. and C.B.; methodology, C.B.; software, C.B.; validation, H.E. and C.B.; formal analysis, C.B.; investigation, H.E.; resources, H.E. and C.B.; data curation, H.E.; writing—original draft preparation, H.E. and C.B.; writing—review and editing, H.E. and C.B.; visualization, C.B.; supervision, C.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset used in this study is available from the corresponding author upon reasonable request after publication of the article.

Acknowledgments: The numerical calculations reported in this paper were fully/partially performed at TUBITAK ULAKBIM, High Performance and Grid Computing Center (TRUBA resources).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Matloob, S.; Li, Y.; Khan, K.Z. Safety Measurements and Risk Assessment of Coal Mining Industry Using Artificial Intelligence and Machine Learning. *Open J. Bus. Manag.* **2021**, *9*, 1198–1209. [CrossRef]
2. Lu, C.; Li, S.; Xu, K.; Zhang, Y. Research on Data-Driven Coal Mine Environmental Safety Risk Assessment System. *Saf. Sci.* **2025**, *183*, 106727. [CrossRef]
3. Strzałkowski, P.; Woźniak, J.; Górniak-Zimroz, J.; Delijewska, B.; Bęś, P.; Solatycka, D.; Janiszewski, M. Identification and Systematics of Safety Hazards in Surface Rock Mining. *Sci. Rep.* **2025**, *15*, 30492. [CrossRef]
4. He, J.; Risso, N.; Bettencourt, T.; Anani, A. Mining the Text: Automating Safety Insights from Mining Accident Reports. In *Proceedings of the 2025 10th International Conference on Machine Learning Technologies, ICMLT 2025*; Institute of Electrical and Electronics Engineers Inc.: New York, NY, USA, 2025; pp. 341–349.
5. Chen, X. Integrated Multimethod Analysis of Miners' Safety Behavior and Risk Interaction for Practical Applications. *Sci. Rep.* **2025**, *15*, 34722. [CrossRef]
6. Xing, Y.; Wu, Y.; Zhang, S.; Wang, L.; Cui, H.; Jia, B.; Wang, H. Discovering Latent Themes in Aviation Safety Reports Using Text Mining and Network Analytics. *Int. J. Transp. Sci. Technol.* **2024**, *16*, 292–316. [CrossRef]
7. Bianchi, F.; Suzgun, M.; Attanasio, G.; Röttger, P.; Jurafsky, D.; Hashimoto, T.; Zou, J. Safety-Tuned LLaMAs: Lessons from Improving the Safety of Large Language Models That Follow Instructions. *arXiv* **2024**, arXiv:2309.07875.
8. Challa, V.; Bright, A.O. Low-Resource Fine-Tuning of LLMs for Domain-Specific Tasks. *Univers. Res. Rep.* **2025**, *12*, 45–56. [CrossRef]

9. Chatterjee, S.; Kadrolli, P.; Kaunda, R.; Miller, H.; Majdara, A. Risk Factors Identification and Injury Severity Classification in Alaska's Mining Industry Using Statistical and Machine Learning Approaches. *Int. J. Min. Reclam. Environ.* **2025**, *39*, 623–640. [CrossRef]
10. Nii-Okai, E. Forecast of Mining Fatalities Using Machine Learning Algorithms. *Preprint* **2025**. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=6045294 (accessed on 23 April 2026).
11. Pishgar, M.; Issa, S.F.; Sietsema, M.; Pratap, P.; Darabi, H. Redeca: A Novel Framework to Review Artificial Intelligence and Its Applications in Occupational Safety and Health. *Int. J. Environ. Res. Public Health* **2021**, *18*, 6705. [PubMed]
12. Donahue, E. Instruction Tuning for Multi-Domain Dialogue Generation in LLMs. *Trans. Comput. Sci. Methods* **2025**, *5*. [CrossRef]
13. Han, X.; Yang, J.; Wang, T.; Bi, Z.; Song, X.; Hao, J.; Song, J. Towards Alignment-Centric Paradigm: A Survey of Instruction Tuning in Large Language Models. *arXiv* **2025**, arXiv:2508.17184. [CrossRef]
14. Xie, T.; Wan, Y.; Huang, W.; Yin, Z.; Liu, Y.; Wang, S.; Linghu, Q.; Kit, C.; Grazian, C.; Zhang, W.; et al. DARWIN Series: Domain Specific Large Language Models for Natural Science. *arXiv* **2023**, arXiv:2308.13565. [CrossRef]
15. Akatsuka, S.; Kumar, A.; Yeow Lee, X.; Vidyaratne, L.; Ghosh, D.; Farahat, A. Rule-Guided Language Model Alignment for Text Generation Management in Industrial Use Cases. In Proceedings of the Neurips Safe Generative AI Workshop 2024, Vancouver, BC, Canada, 14–15 December 2024.
16. Balaskas, G.; Papadopoulos, H.; Pappa, D.; Loisel, Q.; Chastin, S. A Framework for Domain-Specific Dataset Creation and Adaptation of Large Language Models. *Computers* **2025**, *14*, 172. [CrossRef]
17. Zhang, W.; Xiao, S.; Lei, X.; Wang, N.; Zhang, H.; An, M.; Yang, B.; Liu, Z.; Wang, K.; Lian, S. Methodology of Adapting Large English Language Models for Specific Cultural Contexts. *arXiv* **2024**, arXiv:2406.18192. [CrossRef]
18. Chen, J.; Tam, D.; Raffel, C.; Bansal, M.; Yang, D. An Empirical Survey of Data Augmentation for Limited Data Learning in NLP. *Trans. Assoc. Comput. Linguist.* **2023**, *11*, 191–211. [CrossRef]
19. Kiran Evuru, C.; Kumar, S.; Tyagi, U.; Manocha, D. CoDa: Constrained Generation Based Data Augmentation for Low-Resource NLP Sreyan Ghosh. In Proceedings of the Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico, 16–21 June 2024.
20. Yu, J.; Wang, X.; Chen, W. Reliable Data Generation and Selection for Low-Resource Relation Extraction. *Proc. AAAI Conf. Artif. Intell.* **2024**, *38*, 19440–19448. [CrossRef]
21. Chai, Y.; Li, Z.; Liu, J.; Chen, L.; Li, F.; Ji, D.; Teng, C. Compositional Generalization for Multi-Label Text Classification: A Data-Augmentation Approach. *Proc. AAAI Conf. Artif. Intell.* **2024**, *38*, 17727–17735. [CrossRef]
22. Mahamud, M.; Lee, Z.; Samsten, I. Distributional Data Augmentation Methods for Low Resource Language. *arXiv* **2022**, arXiv:2309.04862.
23. Dizlek, O.A.; Yıldız, Z. Evaluation of Received Occupational Safety Measures in Coal Mines. *Karaelmas J. Occup. Health Saf.* **2022**, *6*, 77–86. (In Turkish) [CrossRef]
24. Shaykhlislamova, E.R.; Karimova, L.K.; Beigul, N.A.; Muldasheva, N.A.; Fagamova, A.Z.; Shapoval, I.V.; Volgareva, A.D.; Larionova, E.A. Occupational Health Risk for Workers from Basic Occupational Groups Employed at Copper and Zinc Ore Mining Enterprises: Assessment and Management. *Health Risk Anal.* **2022**, *2*, 107–118. [CrossRef]
25. Keskin, M.Ö.; Doğan, O.; Ersoy, S. Risk Assessment in a Metallic Underground Mining Enterprise, Ore Extraction, Production and Transport Stages. *J. Gaziosmanpasa Sci. Res.* **2020**, *9*, 84–98. (In Turkish)
26. Skripnik, I.; Savelev, D.; Kaverzneva, T.; Rumyantseva, N. Implementation of a Risk-Based OHS Management System at IMC Mining Company. In *Proceedings of the E3S Web of Conferences*; EDP Sciences: London, UK, 2023; Volume 376, p. 05031.
27. Sharma, A.; Kumar, A.; Vardhan, H.; Mangalpady, A.; Mandal, B.B.; Senapati, A.; Avchar, A.; Saini, S. Human-in-the-Loop Data Analytics for Classifying Fatal Mining Accident Causes Using Natural Language Processing and Machine Learning Techniques. *Min. Metall. Explor.* **2025**, *42*, 4155–4167. [CrossRef]
28. Sammour, F.; Xu, J.; Wang, X.; Hu, M.; Zhang, Z. Responsible AI in Construction Safety: Systematic Evaluation of Large Language Models and Prompt Engineering. *J. Constr. Eng. Manag.* **2026**, *152*, 04025217. [CrossRef]
29. Bernardi, M.L.; Cimitile, M.; Panella, G.; Pecori, R.; Simoncelli, G. Automatic Generation of Job Safety Reports with Explainable RAG-Based LLMs. *Inf. Syst. Front.* **2025**, 1–15. [CrossRef]
30. Joeres, R.; Blumenthal, D.B.; Kalinina, O.V. Data Splitting to Avoid Information Leakage with DataSAIL. *Nat. Commun.* **2025**, *16*, 3337. [CrossRef]
31. Apicella, A.; Isgrò, F.; Prevete, R. Don't Push the Button! Exploring Data Leakage Risks in Machine Learning and Transfer Learning. *Artif. Intell. Rev.* **2025**, *58*, 339. [CrossRef]
32. Madrueño, N.; Fernández-Isabel, A.; Cuesta, M.; Lancho, C.; Polo Vera, G.; Martín de Diego, I. Novel Utterance Data Augmentation for Intent Classification Using Large Language Models. *Neural Comput. Appl.* **2025**, *37*, 26711–26736. [CrossRef]
33. Jiang, A.Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D.S.; de las Casas, D.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; et al. Mistral 7B. *arXiv* **2023**, arXiv:2310.06825. [CrossRef]

34. Seller, L.C.; Torres, Í.S.; Vogel-Fernández, A.; Carballo, C.G.; Sánchez, P.M.S.; Martín, A.C.; Ambite, E.d.M. Evaluating Compact LLMs for Zero-Shot Iberian Language Tasks on End-User Devices. *arXiv* **2025**, arXiv:2504.03312. [CrossRef]
35. Zhang, Y. Improving Automatic Clinical Decision Support System with Advanced Computational Methods. Ph.D. Thesis, University of Michigan, Ann Arbor, MI, USA, 2025.
36. Dettmers, T.; Pagnoni, A.; Holtzman, A.; Zettlemoyer, L. QLORA: Efficient Finetuning of Quantized LLMs. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*; Curran Associates Inc.: Red Hook, NY, USA, 2023; pp. 10088–10115.
37. Wiher, G.; Meister, C.; Cotterell, R. On Decoding Strategies for Neural Text Generators. *Trans. Assoc. Comput. Linguist.* **2022**, *10*, 997–1012. [CrossRef]
38. Keneshloo, Y.; Shi, T.; Ramakrishnan, N.; Reddy, C.K. Deep Reinforcement Learning for Sequence-to-Sequence Models. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *31*, 2469–2489. [CrossRef]
39. Dell'anna, D.; Aydemir, F.B.; Dalpiaz, F. Evaluating Classifiers in SE Research: The ECSER Pipeline and Two Replication Studies. *Empir. Softw. Eng.* **2022**, *28*, 3. [CrossRef]
40. Rainio, O.; Teuho, J.; Klén, R. Evaluation Metrics and Statistical Tests for Machine Learning. *Sci. Rep.* **2024**, *14*, 6086. [CrossRef] [PubMed]
41. Lu, N.; Liu, S.; Wu, J.; Chen, W.; Zhang, Z.; Ong, Y.-S.; Wang, Q.; Tang, K. Safe Delta: Consistently Preserving Safety When Fine-Tuning LLMs on Diverse Datasets. *arXiv* **2025**, arXiv:2505.12038. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.