# ABEX-RAT: Synergizing Abstractive Augmentation and Adversarial Training for Classification of Occupational Accident Reports

Jian Chen*,†,#, Jiabao Dou‡,#
*Ningxia Transport Science Research Institute Co., Ltd., Yinchuan, China
†Ningxia Highway Engineering Quality Inspection Center (Co., Ltd.), Yinchuan, China
‡Department of Computer Science, Hong Kong Baptist University, Hong Kong

***Abstract*—The automatic classification of occupational accident reports is pivotal for workplace safety analysis but is persistently hindered by severe class imbalance and data scarcity. In this paper, we propose `ABEX-RAT`, a resource-efficient framework that synergizes generative data augmentation with robust adversarial learning. Unlike computationally expensive large language models (LLMs) fine-tuning, our approach employs a two-stage abstractive-expansive (ABEX) pipeline: it first utilizes a prompt-guided LLM to distill label-critical semantics into concise abstracts, which are then expanded into diverse synthetic samples to balance the data distribution. Subsequently, we train a lightweight classifier using a random adversarial training (RAT) protocol, which stochastically injects perturbations to enhance generalization without significant computational overhead. Experimental results on the OSHA dataset demonstrate that ABEX-RAT establishes a new state-of-the-art, achieving a Macro-F1 score of 90.32% and significantly outperforming both traditional baselines and fine-tuned large models. This confirms that targeted augmentation combined with robust training offers a superior, data-efficient alternative for specialized domain classification.**

***Index Terms*—Occupational Safety, Text Classification, Data Augmentation, Large Language Models, Adversarial Training.**

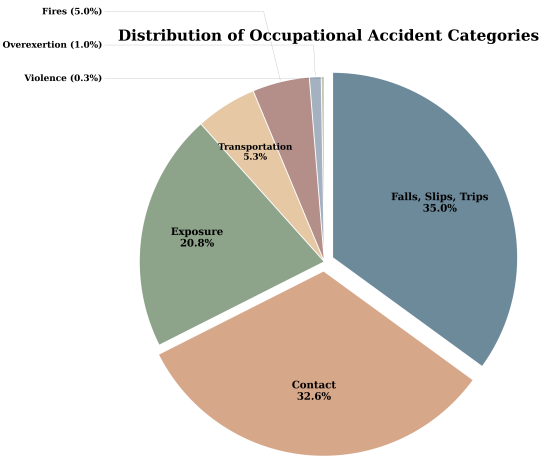


Fig. 1. The problem of category imbalance in occupational accident reports: a case study of OSHA dataset.

#These authors contributed equally to this work. *Corresponding author: jchen.ai@foxmail.com.

## I. Introduction

The analysis of occupational accident reports is a cornerstone of workplace safety management, providing critical insights that help prevent future incidents [1]–[3]. Automatic classification of these free-text narratives into predefined categories is a crucial task, as it enables large-scale statistical analysis and the identification of risk patterns. However, this task is fraught with challenges, chief among them being the severe class imbalance inherent in real-world data [4], as illustrated in Figure 1. Catastrophic but rare events (e.g., *Fires and Explosions*) are vastly outnumbered by more common accident types, leading to models that are biased towards the majority classes and exhibit poor predictive performance on the very incidents that may require the most urgent attention [5]. While recent works have explored LLM-based agents for safety monitoring [6], stable classification of rare accident types remains an open challenge due to the lack of specialized training data.

Prior research has predominantly focused on two paradigms to address this challenge: model-centric adaptation and data-centric augmentation. In the model-centric domain, researchers have attempted to fine-tune pre-trained language models (PLMs) such as BERT and RoBERTa [7], [8]. While these models capture rich semantic features, they often struggle to generalize when faced with extreme data scarcity for minority classes [9]. To mitigate this, cost-sensitive learning strategies, such as the focal loss [10] or re-sampling techniques like SMOTE [11], have been employed to assign higher weights to hard-to-classify examples. However, re-weighting schemes alone cannot compensate for the fundamental lack of feature diversity in underrepresented classes [12].

On the data-centric front, traditional data augmentation techniques have been widely used [13], [14]. Yet, these methods often produce low-quality samples that lack semantic coherence. More recently, the advent of large language models (LLMs) [15], [16] has opened new avenues for generative data augmentation. LLMs demonstrate remarkable capability in generating fluent and diverse text. However, our empirical findings indicate that zero-shot performance of LLMs is often suboptimal for specialized domains like occupational safety due to the lack of domain-specific knowledge [17]. Further-

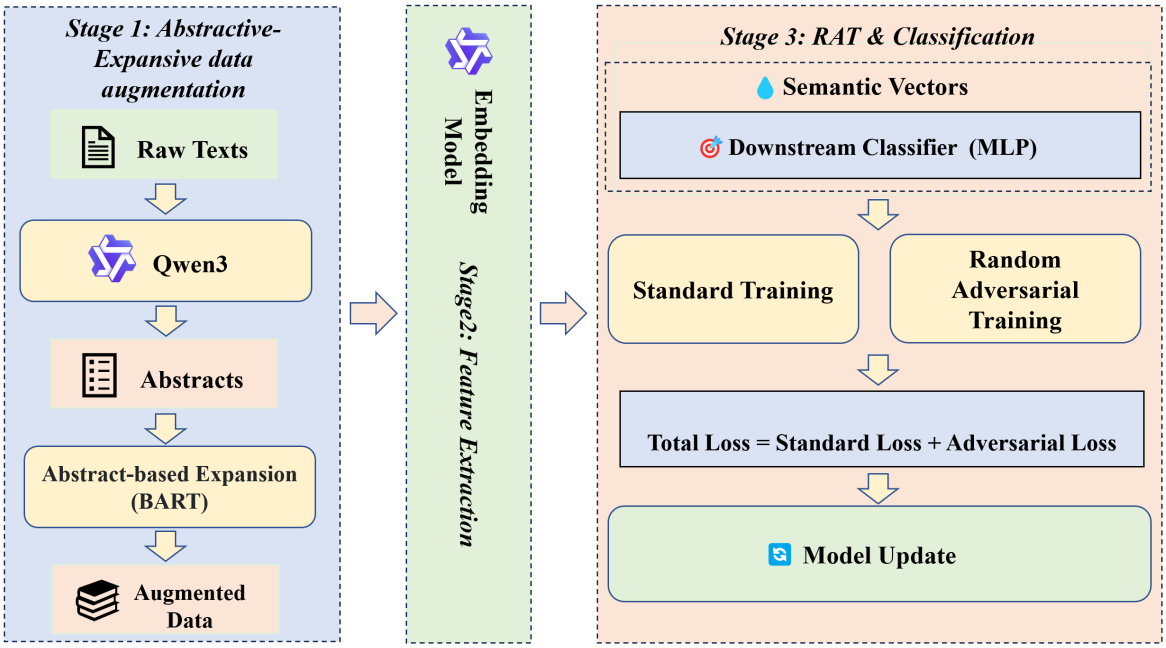


Fig. 2. The overall architecture of our proposed ABEX-RAT framework. The framework consists of three stages. **Stage 1 (ABEX Data Augmentation):** A large language model (e.g., Qwen3) generates a concise abstract from each raw text, which is then expanded into multiple augmented samples by a BART-based model. **Stage 2 (Feature Extraction):** A pre-trained embedding model converts all texts into dense semantic vectors. **Stage 3 (RAT & Classification):** A lightweight MLP classifier is trained on these vectors using a total loss that stochastically combines a standard loss with an adversarial loss to improve model robustness.

more, brute-force fine-tuning of these massive models for data generation is computationally prohibitive for many industrial applications [18].

Parallel to data augmentation, adversarial training (AT) has emerged as a powerful regularization technique to enhance model robustness and generalization [19]. Methods like the fast gradient method (FGM) [20] and projected gradient descent (PGD) [21] operate by introducing small perturbations to the input embeddings, forcing the model to learn smoother decision boundaries. While effective, standard AT methods typically require multiple backward passes to generate perturbations, significantly increasing the training time cost [22]. Consequently, integrating AT into large-scale text classification tasks remains a challenge, particularly when efficiency is a priority.

To address these interconnected limitations—data scarcity, semantic diversity, and training efficiency—we propose `ABEX-RAT`, a novel framework that synergizes a sophisticated data augmentation pipeline with a robust training strategy. Instead of relying on costly full-parameter fine-tuning of LLMs, our approach first enriches the dataset using an abstractive-expansive (ABEX) data augmentation process, inspired by recent advances in generative models [23]. By leveraging a prompt-guided LLM to distill core semantics and a generative model to expand them [24], ABEX generates diverse, high-quality synthetic samples specifically for minority classes.

Subsequently, to handle the augmented feature space robustly, we train a lightweight classifier using a computationally efficient random adversarial training (RAT) protocol [22]. Unlike standard AT, RAT applies perturbations stochastically, thereby enhancing model robustness and generalization without imposing significant computational overhead. This synergistic approach effectively bridges the gap between the need for diverse training data and efficient, robust model optimization.

The main contributions of this work are summarized as follows:

- We propose `ABEX-RAT`, a novel framework that synergizes two components: the ABEX pipeline for generative data augmentation to mitigate class imbalance, and the RAT protocol for computationally efficient adversarial training to enhance model robustness.
- We introduce a specialized prompt-based abstraction and expansion mechanism that ensures the semantic consistency of synthetic accident reports, addressing the hallucinations common in direct LLM generation.
- We demonstrate that on the public OSHA dataset, our method establishes a new state-of-the-art (SOTA), reaching a Macro-F1 score of 90.32% and significantly outperforming traditional fine-tuned models, zero-shot LLMs, and previous SOTA approaches.

## II. METHODOLOGY

To address the challenges of classifying occupational accident reports—specifically the dual issues of data scarcity in critical categories and severe class imbalance—we propose the ABEX-RAT framework. This framework synergizes a generative data augmentation pipeline with a computation-

**Prompt Template for Abstraction Stage**

**# Instruction:**
You are an expert in data augmentation. Convert the provided "Original Incident Report" into a short "Abstract Description" that must: (1) Retain the core event and intent; (2) Remove non-essential details (names, dates, locations) unless critical; (3) Crucially preserve the core meaning of the "Key Concepts" vital for category classification.

**# Constraints:**
**Category:** {category} | **Key Concepts:** {keywords} | **Style:** Concise, factual summary.

**# Task Execution:**
**Input:** `"""` `{full_text}` `"""` → **Output:** [Abstract Description]

Fig. 3. The specific prompt design used in the ABEX abstraction phase. Placeholders are dynamically filled.

efficient robust training strategy. The overall architecture, illustrated in Figure 2, operates as a three-stage pipeline: (1) abstractive-expansive data augmentation to balance the dataset; (2) embedding-based feature extraction; and (3) random adversarial training for robust classification.

### A. Stage 1: ABEX Data Augmentation

The core hypothesis of ABEX is that direct augmentation of raw text often introduces noise or fails to capture the causal logic of accidents. Therefore, we decompose the augmentation process into two sequential steps: *abstraction* and *expansion*.

*1) Prompt-Guided Abstraction:* First, each raw accident report, $T_{\text{raw}}$, is distilled into a concise abstract, $T_{\text{abs}}$, using a LLM. We utilize Qwen3-Instruct[1] as the backbone. Unlike generic summarization, our abstraction function $\mathcal{A}$ is guided by a domain-specific prompt $\mathcal{P}$ designed to retain only the label-defining semantics (e.g., hazard source, incident mechanism, and injury type) while discarding irrelevant stylistic variations and potential personally identifiable information.

$$T_{\text{abs}} = \mathcal{A}_{\text{LLM},\mathcal{P}}(T_{\text{raw}}) \tag{1}$$

To ensure high-quality abstraction, we design a structured prompt $\mathcal{P}$ that acts as a strict instruction set for the LLM. As visualized in Figure 3, the prompt enforces constraints on the output style and explicitly requires the preservation of key concepts specific to the accident category. By constraining the LLM to output a structured summary, we standardize the semantic representation across heterogeneous report styles.

*2) Diversity-Driven Expansion:* Next, to generate synthetic samples, we employ a pre-trained generative model[2] [23] as the expansion function $\mathcal{E}$. This model takes the concise abstract $T_{\text{abs}}$ and hallucinates plausible, grammatically diverse full-text variations, effectively reversing the abstraction process.

To explicitly address class imbalance, the number of generated samples is dynamically calculated. Let $N_c$ be the number of original samples in class $c$, and $N_{\max}$ be the count of the majority class. The expansion factor $R_c$ for class $c$ is defined as:

$$R_c = \left\lceil \lambda \cdot \frac{N_{\max}}{N_c} \right\rceil \tag{2}$$

where $\lambda \in (0, 1]$ is a hyperparameter controlling the degree of balancing. This ensures that minority classes receive significantly more synthetic samples:

$$\{T_{\text{aug}}^{(1)}, \ldots, T_{\text{aug}}^{(R_c)}\} = \mathcal{E}_{\text{BART}}(T_{\text{abs}}) \tag{3}$$

### B. Stage 2: Feature Extraction

Following augmentation, we adopt an efficient embedding-based paradigm rather than fine-tuning the entire LLM [25]. We utilize Qwen3-Embedding model [26] as a fixed feature extractor $\Phi$. This model is optimized for semantic retrieval and clustering, making it ideal for separating confusing accident categories. Each text sample $T_i$ is mapped to a dense vector $\mathbf{x}_i$:

$$\mathbf{x}_i = \Phi(T_i) \in \mathbb{R}^d \tag{4}$$

where $d = 4096$. This process yields a dataset of embedding-label pairs, $\mathcal{D}_{\text{emb}} = \{(\mathbf{x}_i, y_i)\}$, which serves as the static input for the classifier. This design freezes the massive parameters of the LLM, significantly reducing the memory footprint during the training of the downstream classifier.

### C. Stage 3: RAT Classification

The final stage trains a lightweight multi-layer perceptron (MLP) classifier, $f(\mathbf{x}; \theta)$, using the Random Adversarial Training (RAT) mechanism.

To improve robustness against minor lexical shifts in safety reports without the heavy computational overhead of standard adversarial training, RAT applies perturbations stochastically. Following the fast gradient method (FGM) [20], we approximate worst-case perturbations $\mathbf{r}_{\text{adv}}$ bounded by magnitude $\epsilon$. However, rather than computing gradients at every step, a Bernoulli trial $k \sim \text{Bernoulli}(p_{\text{rat}})$ determines whether the adversarial loss is added for each batch:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{std}} + k \cdot \mathcal{L}_{\text{adv}} \tag{5}$$

where $\mathcal{L}_{\text{std}}$ and $\mathcal{L}_{\text{adv}}$ are the losses on clean and perturbed embeddings, respectively. This stochasticity acts as a robust regularizer, preventing the model from overfitting to the adversarial examples themselves.

Furthermore, to directly combat the severe long-tail distribution, we instantiate the base loss function $\mathcal{L}$ with the well-established focal loss [10] rather than standard Cross-Entropy. By utilizing a modulating factor $\gamma$ to down-weight well-classified easy examples, the focal-enhanced RAT objective seamlessly handles both input-level perturbations and dataset-level imbalance.

[1]https://www.modelscope.cn/models/Qwen/Qwen3-235B-A22B-Instruct-2507

[2]https://huggingface.co/utkarsh4430/ABEX-abstract-expand

TABLE I
RESULTS COMPARISON WITH DIFFERENT METHODS. **BOLDED** VALUES INDICATE THE BEST PERFORMANCE.

| Models | Weighted-F1 (%) | | | Macro-F1 (%) | | |
|---|---|---|---|---|---|---|
| | Precision | Recall | **Weighted-F1** | Precision | Recall | **Macro-F1** |
| *Traditional Methods* | | | | | | |
| FastText [27] | 88.62 | 88.59 | 88.59 | 68.29 | 68.52 | 68.39 |
| BERT-base-uncased [7] | 91.48 | 91.50 | 91.47 | 88.97 | 87.93 | 88.43 |
| RoBERTa-base [8] | 90.90 | 90.83 | 90.82 | 91.92 | 88.11 | 89.74 |
| Llama3.1-8B-Instruct [28] | 72.67 | 68.90 | 68.93 | 49.83 | 49.63 | 46.58 |
| Ministral-8B-Instruct [29] | 70.30 | 67.11 | 64.71 | 60.02 | 38.91 | 42.49 |
| Qwen3-8B [30] | 78.87 | 76.51 | 75.75 | 56.91 | 49.05 | 50.88 |
| *Large Language Models for Reasoning* | | | | | | |
| DeepSeek-R1-Distill-Llama-8B [31] | 77.03 | 50.56 | 47.17 | 69.41 | 40.81 | 40.22 |
| Qwen3-8B-think [30] | 82.73 | 77.40 | 74.13 | 72.91 | 60.52 | 61.65 |
| DeepSeek-R1-0528-Qwen3-8B [31] | 83.00 | 77.40 | 77.07 | 64.03 | 66.90 | 58.50 |
| *SOTA Models* | | | | | | |
| $Qwen3-8B_{SFT}$ | 81.04 | 76.06 | 74.59 | 68.76 | 54.99 | 58.67 |
| Qwen3-Embedding-8B [26] | 91.85 | 91.95 | 91.76 | 73.86 | 69.96 | 71.64 |
| INSTRUCTOR-CIT [4] | 88.83 | 88.81 | 88.77 | 84.24 | 82.52 | 83.14 |
| ABEX-RAT (Ours) | 93.07 | 92.84 | **92.82** | 89.99 | 91.88 | **90.32** |

## III. EXPERIMENTS

### A. Experimental Setup

**Dataset and Metrics:** We evaluate our framework on the OSHA dataset[3] [3], containing 4,770 construction accident reports across seven categories. Data is partitioned into a stratified 8:1:1 split for training, validation, and testing to preserve the real-world class imbalance. We report Weighted-F1 for overall accuracy and Macro-F1 to rigorously evaluate performance on the critical, long-tail incident types.

**Implementation Details:** Models are implemented in PyTorch and trained on a single NVIDIA RTX 4090 GPU. The downstream classifier is a 2-layer MLP with ReLU. Hyperparameters are set as follows: learning rate $\eta = 10^{-4}$, batch size $B = 16$, and epochs $E = 100$. For the RAT module, we set $p_{\text{rat}} = 0.5$ and $\epsilon = 0.1$. The focal loss uses $\gamma = 3.0$.

**Baselines:** We benchmark `ABEX-RAT` against three method categories: (1) *Traditional Fine-tuning:* `FastText`, `BERT`, and `RoBERTa` trained on raw data; (2) *Zero-shot LLMs:* `Llama3.1-8B-Instruct` and `Qwen3-8B` to evaluate intrinsic domain knowledge without parameter updates; and (3) *Domain-Specific SOTA:* `INSTRUCTOR-CIT` and a LoRA-fine-tuned $\texttt{Qwen3-8B}_{\texttt{SFT}}$ [32], representing brute-force LLM adaptation.

### B. Main Results and Analysis

The quantitative comparison results are summarized in Table I. Our proposed `ABEX-RAT` framework achieves a new state-of-the-art performance on the OSHA dataset, recording a Weighted-F1 of 92.82% and a Macro-F1 of 90.32%. These metrics confirm that our synergistic approach robustly handles the dichotomy between frequent and rare accident categories.

[3]https://github.com/qiao77/Injury-Narratives

`ABEX-RAT` outperforms the strongest traditional baseline, `RoBERTa-base`, by 2.0% in Weighted-F1 and 0.58% in Macro-F1. While the numerical margin in Macro-F1 appears modest, it is critical to note that our method achieves this using a lightweight MLP on fixed embeddings, avoiding the heavy computational inference cost associated with full Transformer models. The zero-shot performance of general-purpose LLMs (e.g., `Qwen3-8B`) is substantially lower, lagging behind supervised methods by over 40% in Macro-F1. This underscores that despite their linguistic capabilities, generalist LLMs lack the specific taxonomic knowledge required for precise occupational safety classification, validating the need for task-specific supervision.

A pivotal finding is that `ABEX-RAT` surpasses the fine-tuned $\texttt{Qwen3-8B}_{\texttt{SFT}}$ by a remarkable margin. This result suggests that on small, imbalanced datasets, brute-force fine-tuning of billions of parameters leads to severe overfitting on majority classes. In contrast, our strategy of freezing the LLM backbone + augmenting the data space + smoothing the decision boundary proves to be a far more effective and data-efficient paradigm. Furthermore, the 18.68% jump in Macro-F1 over the `Qwen3-Embedding` baseline explicitly isolates the contribution of our contributions, proving that the high performance stems from the proposed augmentation and adversarial training rather than merely the quality of the base embeddings.

## IV. DISCUSSION

### A. Ablation Study: Decoupling Synergy

To evaluate individual component contributions, we compared `ABEX-RAT` against two variants: `w/o RAT` (augmented data with standard cross-entropy) and `w/o ABEX` (RAT on original imbalanced data). As shown in Figure 4, `w/o RAT` achieves an 86.68% Macro-F1, proving ABEX effectively

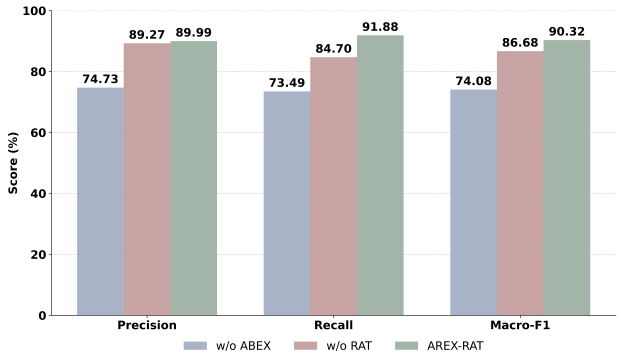


Fig. 4. Results of the ablation experiment.

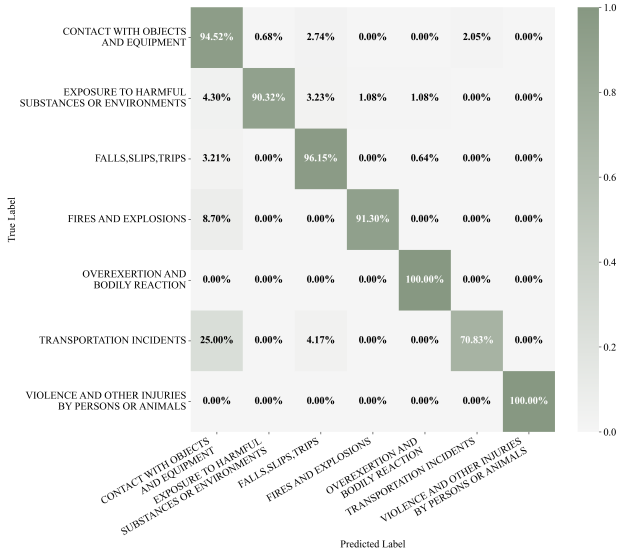


Fig. 5. Normalized confusion matrix.

injects semantic diversity into minority classes, though it risks overfitting synthetic patterns. Conversely, `w/o ABEX` yields only 74.08% Macro-F1, highlighting that adversarial training cannot invent missing knowledge on sparsely populated data manifolds. `ABEX-RAT` outperforms both with a 90.32% Macro-F1. This synergy is particularly evident in Recall (jumping from 84.70% to 91.88%), suggesting RAT effectively smooths decision boundaries around ABEX's synthetic samples, significantly improving generalization on challenging minority classes.

### B. Fine-grained Performance Analysis

To further dissect classification performance, we analyze the normalized confusion matrix (Figure 5). The diagonal elements demonstrate robust recall across all categories, specifically distinguishing the framework's efficacy on the rarest class, *FIRES AND EXPLOSIONS* (91.30%). This is a critical improvement over baseline methods which often ignore this category entirely due to sample scarcity. The analysis also reveals semantically plausible misclassifications. For instance, 25% of *TRANSPORTATION INCIDENTS* are predicted as

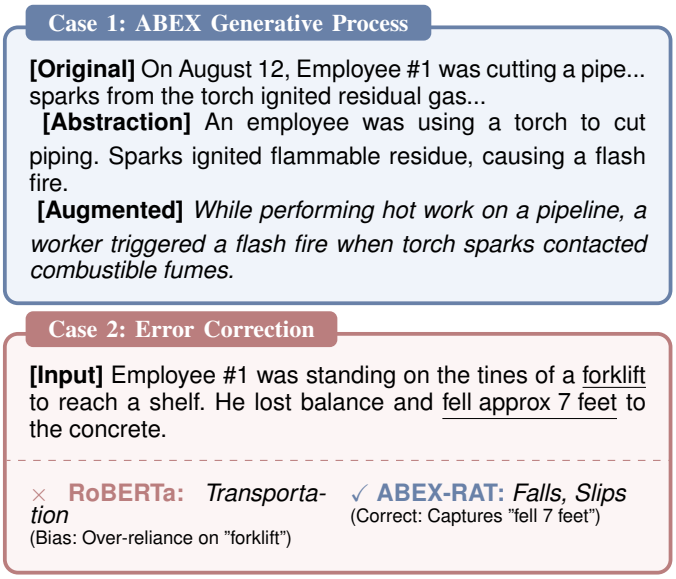


Fig. 6. Case Study of the proposed framework. **Top:** Semantic consistency of ABEX generation. **Bottom:** Robustness in correcting hard samples where baselines fail due to keyword bias.

*CONTACT WITH OBJECTS*. A manual inspection reveals these often involve vehicles striking pedestrians, which sits at the semantic intersection of both categories. This suggests that future work could benefit from multi-label modeling or hierarchical classification to resolve such ambiguities.

### C. Case Study

To intuitively understand the improvements, we visualize real-world examples in Figure 6.

**Case 1:** The upper panel demonstrates the ABEX pipeline. The abstractive step successfully strips dates, names while retaining the causal chain. The expansion step then generates a grammatically distinct variation that preserves these semantics, effectively enriching the feature space for the *Fire* category.

**Case 2:** The lower panel illustrates a hard sample involving a forklift. The baseline `RoBERTa` model, likely biased by the high-frequency keyword forklift, incorrectly predicts *TRANSPORTATION INCIDENT*. However, `ABEX-RAT` correctly identifies the root cause as *FALLS, SLIPS, TRIPS*. This indicates that our model has learned to attend to the mechanism of injury rather than just surface-level nouns, validating the robustness gain from adversarial training.

## V. Conclusion

In this paper, we presented ABEX-RAT, a data-efficient framework tailored for the high-stakes task of occupational accident report classification. By synergizing an LLM-driven abstractive-expansive data augmentation pipeline with a lightweight RAT protocol, we effectively mitigated the dual challenges of extreme data scarcity and class imbalance. Empirical evaluations on the OSHA dataset confirm that ABEX-RAT establishes a new state-of-the-art, achieving a Macro-F1 score of 90.32% and significantly outperforming both traditional baselines and computationally expensive large model fine-tuning. These findings validate that targeted

data enrichment combined with robust regularization offers a superior, resource-efficient alternative to brute-force model scaling for specialized domain applications. Future research will extend this paradigm to multi-label scenarios and explore hierarchical classification techniques to better resolve semantic overlaps between closely related accident categories.

## ACKNOWLEDGMENT

This study was supported by the Transportation Power Special Fund Project of the Ningxia Hui Autonomous Region (No. 202500210) and the Ningxia Natural Science Foundation (No. 2026AAC020063).

## REFERENCES

[1] Antoine J-P Tixier, Matthew R Hallowell, Balaji Rajagopalan, and Dean Bowman. Automated content analysis for construction safety: A natural language processing system to extract precursors and outcomes from unstructured injury reports. *Automation in Construction*, 62:45–56, 2016.

[2] Fan Zhang, Hasan Fleyeh, Xinru Wang, and Minghui Lu. Construction site accident analysis using text mining and natural language processing techniques. *Automation in Construction*, 99:238–248, 2019.

[3] Jianfeng Qiao, Changfeng Wang, Shuang Guan, and Lv Shuran. Construction-accident narrative classification using shallow and deep learning. *Journal of Construction Engineering and Management*, 148(9):04022088, 2022.

[4] Qing Shuang, Xishan Liu, Zhaojing Wang, and Xinxin Xu. Automatically categorizing construction accident narratives using the deep-learning model with a class-imbalance treatment technique. *Journal of Construction Engineering and Management*, 150(9):04024107, 2024.

[5] May Shayboun, Dimosthenis Kifokeris, and Christian Koch. A review of machine learning for analysing accident reports in the construction industry. *Journal of Information Technology in Construction*, 30:439–460, 2025.

[6] Wei Li, Fuqi Ma, Zhiyuan Zuo, Rong Jia, Bo Wang, and Abdullah M Alharbi. Safetygpt: An autonomous agent of electrical safety risks for monitoring workers' unsafe behaviors. *International Journal of Electrical Power & Energy Systems*, 168:110672, 2025.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

[8] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[9] Rana Khallaf and Mohamed Khallaf. Classification and analysis of deep learning applications in construction: A systematic literature review. *Automation in construction*, 129:103760, 2021.

[10] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[11] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[12] Jian Chen, Leilei Su, Yihong Li, Mingquan Lin, Yifan Peng, and Cong Sun. A multimodal approach for few-shot biomedical named entity recognition in low-resource languages. *Journal of Biomedical Informatics*, 161:104754, 2025.

[13] Jason Wei and Kai Zou. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*, 2019.

[14] Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 86–96, 2016.

[15] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[16] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[17] Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. The curse of recursion: Training on generated data makes models forget. *arXiv preprint arXiv:2305.17493*, 2023.

[18] Jian Chen, Jinbao Tian, Ting Zheng, Yingxin Hui, and Zhou Li. Qwen-aec: Instruction-tuning large language models for high-performance building code analysis. *Advanced Engineering Informatics*, 71:104268, 2026.

[19] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[20] Takeru Miyato, Andrew M Dai, and Ian Goodfellow. Adversarial training methods for semi-supervised text classification. *arXiv preprint arXiv:1605.07725*, 2016.

[21] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[22] Jian Chen, Shengyi Lv, and Leilei Su. Ranat4bie: Random adversarial training for biomedical information extraction. In *2025 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2025.

[23] Sreyan Ghosh, Utkarsh Tyagi, Sonal Kumar, CK Evuru, S Ramaneswaran, S Sakshi, and Dinesh Manocha. Abex: data augmentation for low-resource nlu via expanding abstract descriptions. *arXiv preprint arXiv:2406.04286*, 2024.

[24] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 7871–7880, 2020.

[25] Jian Chen, Jinbao Tian, Wangdong Xu, Wenbo Xia, and Zhou Li. Weaving embeddings: A fine-tune-free llm fusion framework for efficient health and safety text analysis. In *2025 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 6676–6683. IEEE, 2025.

[26] Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025.

[27] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*, 2016.

[28] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[29] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023.

[30] Qwen Team. Qwen3 technical report, 2025.

[31] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[32] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.