Computer-Aided Civil and Infrastructure Engineering 49 (2026) 100098

Contents lists available at ScienceDirect

Computer-Aided Civil and Infrastructure Engineering

journal homepage: www.sciencedirect.com/journal/computer-aided-civil-and-infrastructure-engineering

An instruction fine-tuning and retrieval-augmented generation framework for intelligent defect diagnosis and maintenance decision support in high-speed railway turnouts☆

Yi Wang [a, b], Xiaopei Cai [a, b,*], Bin Cui [c], Xueyang Tang [a, b], Clare Wood [d], Yue Hou [d,**]

[a] State Key Laboratory of Advanced Rail Autonomous Operation, Beijing Jiaotong University, Beijing, 100044, China
[b] School of Civil Engineering, Beijing Jiaotong University, Beijing, 100044, China
[c] National Engineering Laboratory for Digital Construction and Evaluation of Urban Rail Transit, Tianjin, 300308, China
[d] Faculty of Science and Engineering, Swansea University, Swansea, SA2 8PP, United Kingdom

HIGHLIGHTS

- An LLM-driven system for railway turnout defect diagnosis and maintenance decisions.
- Textualization of data enables semantic understanding by the LLM for diagnosis.
- Enhanced fine-tuning via contrastive loss achieves 89.6% diagnostic accuracy.
- RAG integration automates the generation of standardized maintenance decisions.

ARTICLE INFO

Keywords:
High-speed railway
Turnout
Large language model
Defect diagnosis
Intelligent operation and maintenance

ABSTRACT

High-speed railway (HSR) turnouts are among the most mechanically demanding components in the railway infrastructure, yet current operation and maintenance (O&M) practices remain largely reactive, experience-dependent, and disconnected from automated decision support. This paper presents a large language model (LLM)-driven expert system that bridges the gap between raw sensor data and actionable maintenance decisions for turnout defect diagnosis. Three core contributions are made. First, a data textualization strategy is developed to convert train body acceleration signals into structured text sequences comprehensible to LLMs, enabling domain-specific diagnosis without architectural modification of the base model. Second, an enhanced instruction fine-tuning scheme is proposed, incorporating a contrastive loss function that tightens intra-class feature clusters and widens inter-class margins, alongside a hierarchical evaluation method that reliably extracts categorical intent from free-form model outputs. Third, a retrieval-augmented generation (RAG) module is integrated with the fine-tuned model, enabling the system to generate standards-compliant maintenance recommendations directly from diagnostic results. Controlled experiments across four pre-trained models and 26 experimental groups demonstrate that the proposed system reaches a peak diagnostic accuracy of 89.6%, while preserving the natural language generation capabilities essential for report production. The framework is evaluated on a physically representative dataset generated by a validated stochastic vehicle–turnout dynamics model. The resulting integrated pipeline, from extracted signal features to maintenance decision output, offers a practical and scalable solution for intelligent O&M of complex railway turnout infrastructure and beyond.

☆ This work was supported by the State Key Laboratory of Advanced Rail Autonomous Operation (RAO2025ZT002), Beijing Jiaotong University; Tianjin Key R&D Programme for Beijing–Tianjin–Hebei Collaborative Innovation (25YFXTHZ00260); the Natural Science Foundation of Beijing, China (L251029); the Fundamental Research Funds for the Central Universities (2025QYBS007); and the Key Research Project of China Railway Design Corporation (2024A0253805).

* Corresponding author at: State Key Laboratory of Advanced Rail Autonomous Operation, Beijing Jiaotong University, Beijing, 100044, China.
** Corresponding author.
Email addresses: 24110580@bjtu.edu.cn (Y. Wang), xpcai@bjtu.edu.cn (X. Cai), cuibin@crdc.com (B. Cui), 98940436@bjtu.edu.cn (X. Tang), c.wood@swansea.ac.uk (C. Wood), yue.hou@swansea.ac.uk (Y. Hou).

https://doi.org/10.1016/j.cacaie.2026.100098
Received 29 March 2026; Received in revised form 10 May 2026; Accepted 16 May 2026
Available online 21 May 2026

1. Introduction

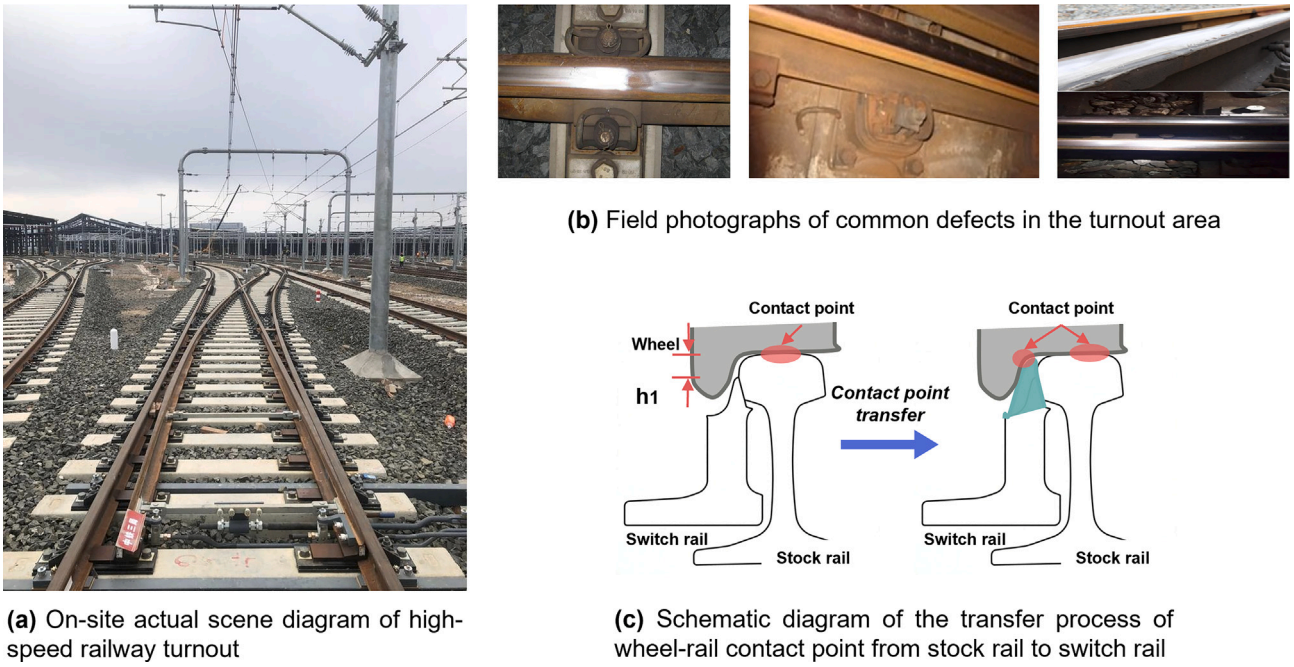
Turnouts are among the most failure-prone components in high-speed railway (HSR) track systems. Their complex geometry and concentrated stress distribution make them susceptible to geometric deviations that, if left unaddressed, can develop into rail corrugation, fractures, or structural instability (Cai et al., 2024; La Paglia et al., 2023). As rail networks continue to expand globally, the volume and diversity of monitoring data generated from these components grow correspondingly. This growth places increasing pressure on operation and maintenance (O&M) workflows, which in many organizations have yet to move beyond manual inspection and experience-based judgment. A key geometric indicator in this context is the reduction value, defined as the vertical height difference between the top surface of the switch or point rail and the adjacent stock rail (Chang et al., 2022). Maintaining this parameter within prescribed tolerances is essential for smooth wheel–rail load transfer as trains negotiate switches. When the reduction value deviates from its design specification, the load transfer path shifts unfavorably: an excessively large value delays wheel load transition, inducing lateral sway and impact forces; conversely, an excessively small value causes premature wheel–switch rail contact, intensifying local stress concentration and accelerating material fatigue (Chen et al., 2020, 2024). Accurate and automated perception of this geometric state therefore constitutes a prerequisite for any credible intelligent maintenance framework (Fig. 1).

Current monitoring approaches fall into two broad categories. Fixed-point microcomputer-based systems can track electromechanical parameters such as switch machine current and switching force in real time, but they struggle to directly capture the structural signatures of reduction deviations, which limits their value for precise defect diagnosis (Wang et al., 2024). Periodic inspections using track geometry cars or manual patrols provide richer acceleration data as trains traverse turnouts (Li et al., 2023, 2014), yet the subsequent analysis pipeline depends heavily on expert interpretation and does not readily scale to the data volumes generated in dense rail networks. In a broader context, structural health monitoring methodologies have progressed from early signal processing techniques (Amezquita-Sanchez & Adeli, 2016) and statistical anomaly detection (Entezami et al., 2025) toward lower-cost, higher-coverage, and increasingly automated platforms (Sarmadi et al., 2023). For turnouts specifically, the remaining challenge is to connect automated data acquisition with integrated maintenance decision generation within a single workflow.

Large language models (LLMs) have emerged as a promising avenue for bridging this gap. Their capacity for sequence understanding, contextual reasoning, and natural language generation offers a unified mechanism for transforming heterogeneous sensor inputs into structured, human-readable outputs (Wang et al., 2024; Zhao et al., 2025). Recent studies have shown their utility in mechanical fault diagnosis (Lin et al., 2025; Tao et al., 2025), civil infrastructure assessment (Jiang et al., 2025; Xu et al., 2025), and transportation safety analysis (Jia et al., 2025; Liu et al., 2025), among other engineering domains. The wider adoption of deep learning techniques in railway infrastructure, including UAV-based inspection (Aela et al., 2024) and deep transfer learning for image-based structural damage recognition (Gao & Mosalam, 2018), provides further evidence of the field's receptiveness to data-driven automation. A detailed review of these recent developments is provided in Section 2.3.

Despite these advances, no existing study has developed an integrated system that simultaneously handles the full pipeline from structured signal features to standardized maintenance recommendations for railway turnout components. This paper addresses that gap directly. The proposed framework, illustrated in Fig. 2, combines instruction fine-tuning of a general-purpose LLM using the LlamaFactory platform (Zheng et al., 2024) with a retrieval-augmented generation (RAG) module grounded in industry maintenance standards (Zhao et al., 2024), constructing an expert system that produces both a defect classification and a professional maintenance report from a single inference pass. While the current implementation relies on predefined time–frequency features, the diagnostic and decision-generation stages operate without manual intervention once the features are extracted.

The remainder of this paper is organized as follows. Section 2 reviews the relevant literature. Section 3 details the proposed methodology. Section 4 describes the experimental setup. Section 5 presents and discusses the results. Section 6 focuses on the main limitations of the current study and outlines directions for future research. Section 7 concludes with a summary of findings.



(a) On-site actual scene diagram of high-speed railway turnout

(b) Field photographs of common defects in the turnout area

(c) Schematic diagram of the transfer process of wheel-rail contact point from stock rail to switch rail

Fig. 1. Illustration of a typical HSR turnout structure.

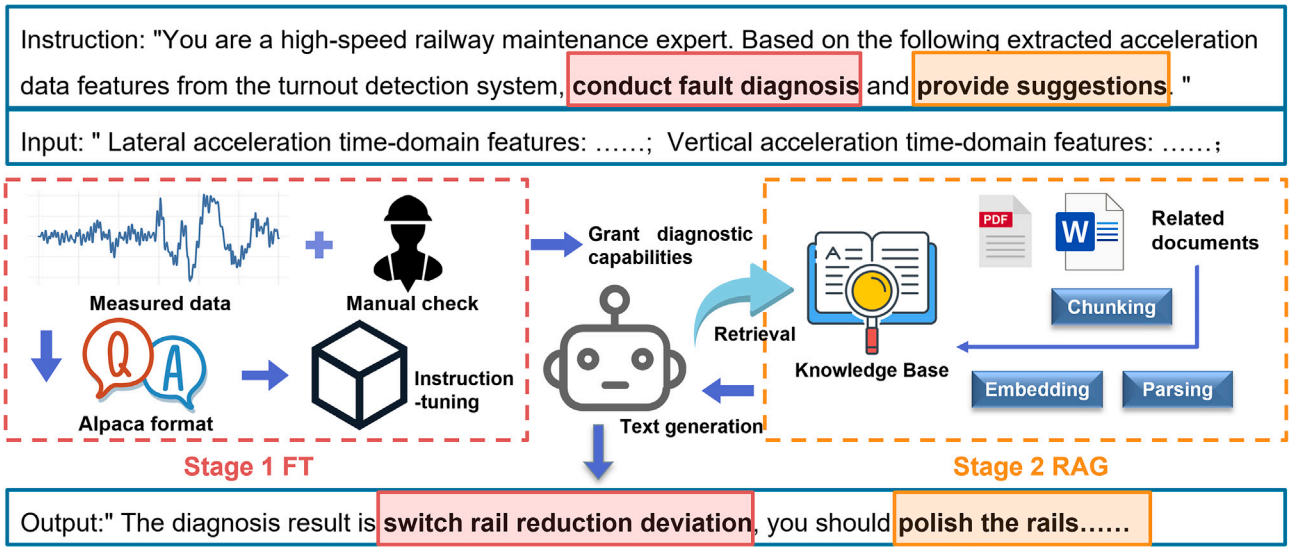


Fig. 2. Schematic diagram of the proposed expert system.

2. Related work

This section reviews the three areas most relevant to the present study: the physical relationship between turnout reduction deviation and measurable vehicle response (Section 2.1), the evolution of maintenance paradigms for HSR infrastructure (Section 2.2), and recent LLM applications in engineering monitoring and diagnosis, organized as a structured gap analysis across four functional dimensions (Section 2.3).

2.1. Correlation between turnout reduction deviation and train body response

Reduction deviations at the wheel–rail interface disrupt smooth contact and generate measurable dynamic responses throughout the vehicle. While conventional safety assessments have concentrated on local indicators—wheel–rail forces and derailment coefficients—that require specialized trackside instrumentation, train body acceleration offers a globally accessible alternative that is routinely collected during commercial operations and contains rich information about the underlying wheel–rail interaction state.

Early investigations relied on multi-body dynamics models, which captured macroscopic vehicle behavior but were limited in their ability to resolve medium-to-high frequency vibration transmission to the car body. Wang et al. (2024) examined the influence of reduction deviation on wheel–rail forces, while Chang et al. (2022) characterized the macroscopic dynamic consequences of suboptimal reduction values. These studies established the qualitative link between reduction value deviation and vehicle response, but did not provide a quantitative basis for automated diagnosis. The introduction of multi-flexible-body simulation by Ma et al. (2025) addressed this limitation by showing that distinct reduction deviation types (whether originating from the switch rail, the point rail, or both simultaneously) excite characteristic time–frequency signatures in train body vibration, thereby laying the physical foundation for inversion-based diagnosis. The stochastic dynamics model employed in the present study, which accounts for uncertainties across 27 vehicle and turnout parameters and uses measured spectra to parametrize random irregularities, extends this line of work by providing a controlled yet physically representative dataset for LLM fine-tuning (Tang et al., 2025).

2.2. Evolution of O&M paradigms for HSR

Turnout maintenance methodology has passed through three identifiable phases, each characterized by a different relationship between data, expertise, and automation.

The conventional paradigm (Fig. 3(a)) couples periodic data collection with manual expert interpretation, following established maintenance standards (Avci et al., 2021; Cao et al., 2020). Diagnostic quality under this paradigm is inherently bounded by the availability and consistency of qualified personnel, and the turnaround time from data acquisition to maintenance action is often measured in days or weeks. The data-driven paradigm (Fig. 3(b)) replaces expert interpretation with deep learning models trained to extract defect-state mappings from large datasets, substantially improving throughput and objectivity (Aela et al., 2024; Deng et al., 2022). However, the interpretability limitations of these black-box models mean that final maintenance decisions still require human oversight, and the models themselves produce categorical labels rather than actionable guidance. The LLM-augmented paradigm (Fig. 3(c)) addresses interpretability and decision generation simultaneously: by pairing existing diagnostic conclusions with RAG-based knowledge retrieval, LLMs can produce standardized, auditable maintenance recommendations without additional fine-tuning (Lee et al., 2024). The neural network foundations that have enabled this progression trace back to early applications of connectionist models in civil engineering (Adeli, 2001; Adeli & Yeh, 1989), which laid the intellectual groundwork for current deep learning and language model deployments.

The present study advances beyond all three paradigms by proposing an integrated expert system (Fig. 3(d)) in which a single fine-tuned LLM handles the workflow from extracted signal features to maintenance report generation within a unified architecture, thereby reducing the dependency on separate diagnostic and decision-support modules.

2.3. LLM applications in infrastructure monitoring and diagnosis

Recent LLM deployments in engineering can be organized along four functional dimensions relevant to the present work: time-series diagnosis, knowledge retrieval, decision generation, and multimodal integration.

In the area of time-series diagnosis, primary efforts have focused on adapting LLMs for fault diagnosis from sensor signals (Lin et al.,

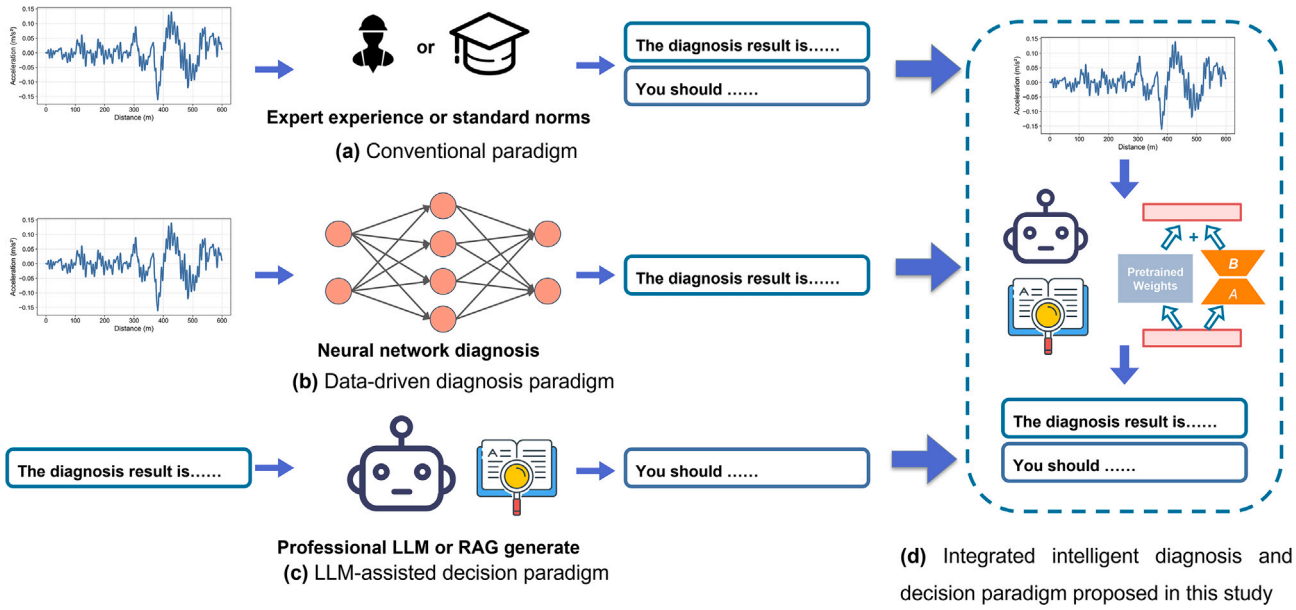


Fig. 3. Schematic diagram of the evolution of maintenance paradigm.

2025; Zhang et al., 2025), typically through prompt-based or fine-tuning approaches that convert numerical sequences into textual representations. More recently, multimodal architectures that fuse vibration spectrograms with task-specific prompts via low-rank adaptation (LoRA)-adapted LLMs have achieved cross-dataset generalization with explainable outputs (Wang et al., 2025). However, these studies predominantly address mechanical systems (bearings, gears) rather than civil infrastructure, and none integrates the diagnostic output with downstream maintenance decision workflows.

Regarding knowledge retrieval, RAG has been applied to construction management documentation (Wu et al., 2025) and safety knowledge retrieval for building projects (Lee et al., 2024). In structural health monitoring, convolutional neural network (CNN)-based damage classifiers have been coupled with LoRA-fine-tuned LLMs to synthesize heterogeneous sensor data and inspection reports into prescriptive maintenance decisions (Smarsly et al., 2025). These systems demonstrate the value of grounding LLM outputs in domain standards, but operate independently of any upstream automated diagnostic model, leaving a gap between data-driven diagnosis and knowledge-grounded recommendations.

In the domain of decision generation, chain-of-thought reasoning guided by the Human Factors Analysis and Classification System (HFACS) taxonomy has been used to analyze aviation accident narratives (Liu et al., 2025). Image-based zero-shot prompting has been applied to pavement condition assessment (Xu et al., 2025). Post-earthquake structural damage reports have been generated from image–text pairs (Gao et al., 2026; Jiang et al., 2025). In railway-specific contexts, graph neural network (GNN)–LLM integration has been explored for track settlement forecasting (Zhou et al., 2026), and LLMs have been used for bridge specification generation (Maharjan & Chun, 2026). While these applications produce actionable outputs, they rely on manually curated inputs rather than automated diagnostic pipelines, and do not incorporate standardized maintenance knowledge through retrieval mechanisms.

For multimodal integration, approaches combining sensor data with text or images have been explored for landslide interpretation (Areerob et al., 2025), mechanical fault analysis (Jose et al., 2024; Li et al., 2025), and knowledge graph construction (Zhou et al., 2024). Physics-informed evaluation frameworks have assessed LLM-generated driving scenarios (Jia et al., 2025), and natural language processing for transportation knowledge extraction has been surveyed more broadly (Zhang et al., 2026). These studies demonstrate the potential of multimodal LLM systems but do not close the loop between sensor-driven diagnosis and standards-compliant decision outputs.

Across these four dimensions, a consistent gap remains: existing systems address either the knowledge retrieval dimension or the diagnostic dimension in isolation, but no study has combined both within an integrated pipeline that operates on structured sensor-derived features and produces maintenance recommendations grounded in industry standards. The present study is designed to bridge this gap.

3. Method

This section describes the proposed expert system in detail. Section 3.1 covers the construction of the fine-tuned defect diagnosis model, including signal feature extraction, the improved training objective, and the task-adapted evaluation strategy. Section 3.2 describes the RAG module for standardized maintenance recommendation generation.

3.1. Construction of the defect diagnosis LLM via instruction fine-tuning

The goal of this stage is to convert a general-purpose LLM into a domain expert capable of classifying turnout reduction value deviation defects from structured textual representations of sensor signals. The overall process is shown in Fig. 4. Rather than modifying the model architecture or training a new model from scratch, the approach relies on parameter-efficient instruction fine-tuning within the LlamaFactory framework (Zheng et al., 2024), supplemented by a contrastive learning objective and a task-adapted evaluation strategy.

3.1.1. Signal feature extraction and textual dataset construction

Raw train body acceleration signals are not directly suitable for LLM processing, both because of their high dimensionality and because the statistical patterns relevant to turnout health are not transparently encoded in the raw time series. A textualization strategy is therefore adopted to transform the signal pattern recognition problem into a language understanding task, making the diagnostic problem accessible to the full range of LLM capabilities, including reasoning, contextual grounding, and natural language report generation.

Drawing on established practice in rotating machinery fault diagnosis (Tao et al., 2025), 24 features are extracted from the lateral and vertical acceleration channels: 12 time-domain quantities and 12 frequency-domain quantities, all carrying clear physical interpretations.

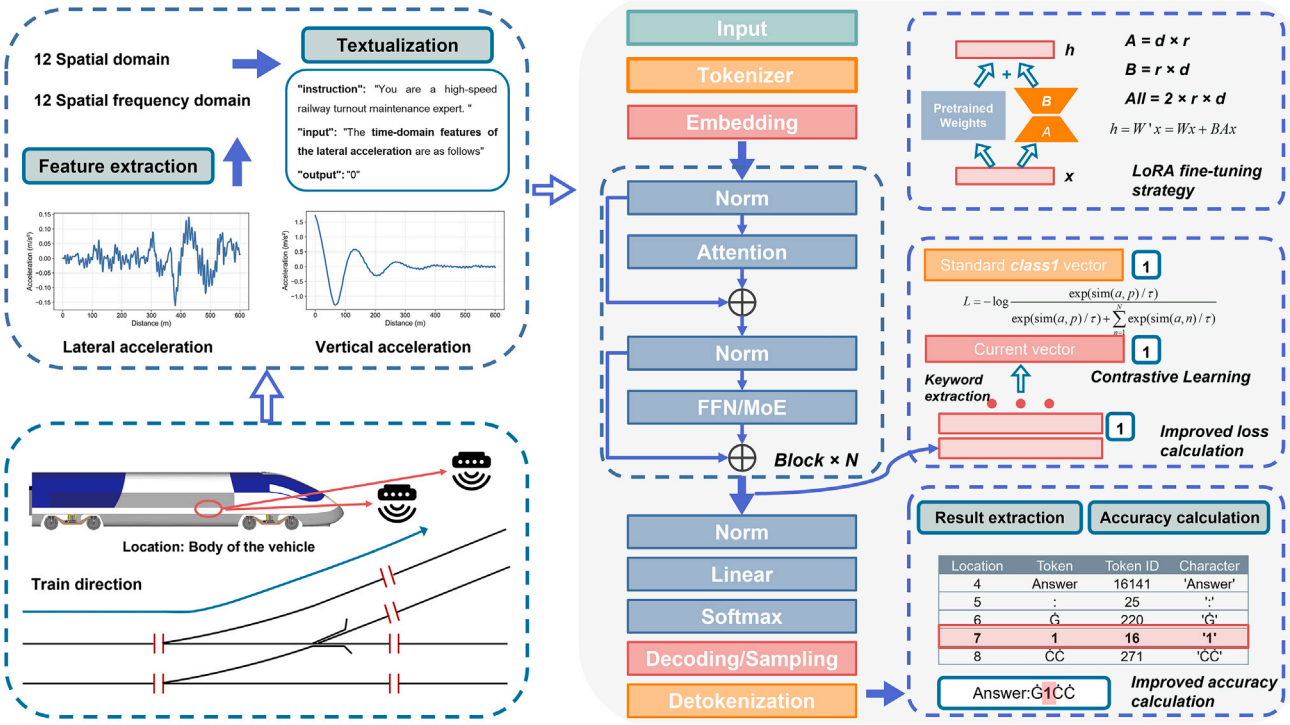


Fig. 4. Schematic diagram of the defect diagnosis LLM based on instruction fine-tuning.

Table 1
Feature parameters of high-speed train body acceleration signals.

Type of Feature	Feature Name
Time-domain	Mean value, Standard deviation, Square root amplitude, Absolute mean value, Peak value, Skewness, Kurtosis, Variance, Kurtosis index, Peak index, Waveform index, Pulse index
Frequency-domain	Frequency mean value, Frequency variance, Frequency skewness, Frequency kurtosis, Gravity frequency, Frequency standard deviation, Frequency root mean square, Average frequency, Regularity degree, Variation parameter, Eighth-order moment, Sixteenth-order moment

These are listed in Table 1. The selection prioritizes features with well-established physical meaning for two reasons: first, physically interpretable features allow the instruction prompt to include meaningful guidance (e.g., directing the model's attention to the most discriminative quantities identified by analysis of variance (ANOVA)), which would not be possible with abstract latent representations; second, transparent features facilitate future extension of the defect taxonomy without requiring redesign of the feature extraction pipeline.

The extracted features are arranged into instruction-following data records comprising three components, following prompt engineering conventions discussed in (Luo et al., 2023). The **Instruction** component specifies the model's role as a domain expert, defines the classification task, prescribes the output format, and provides brief physical guidance on the most discriminative features identified by ANOVA; this guidance directs attention to specific feature names and their expected behavior under each defect condition, giving the model a physically grounded reasoning anchor rather than relying solely on statistical pattern matching. The **Input** component lists each feature name alongside its corresponding lateral and vertical value in plain natural language; by presenting numerical values within named feature contexts (e.g., "Lateral Standard Deviation: 0.0372"), the format allows the LLM to leverage its pre-trained understanding of statistical concepts when interpreting the data. The **Output** component is a single integer label in {0, 1, 2, 3} corresponding to the defect condition; abstract labels are adopted rather than descriptive category names because they accommodate future expansion of the defect taxonomy without restructuring the training data and focus the model's optimization on categorical mapping, thereby reducing sensitivity to vocabulary choices and the ambiguity they may introduce. A representative example of this format is shown in Fig. 5.

3.1.2. Training objective enhancement with contrastive learning

Standard instruction fine-tuning minimizes cross-entropy loss over next-token prediction, which drives the model to produce correct textual outputs but does not explicitly shape the geometry of the learned representations. In a classification setting, this can leave the decision boundaries between defect categories poorly defined, particularly when the feature differences between categories are subtle—as is the case for reduction value deviations, where conditions such as switch rail deviation alone and simultaneous switch-point rail deviation produce partially overlapping acceleration signatures.

To address this, a contrastive learning term is incorporated into the training objective, drawing on the formulation introduced by Hadsell et al. (2006). This mechanism enforces that representations of samples belonging to the same defect category cluster tightly in the embedding space, while representations from different categories are pushed apart. Implementation details are illustrated in the "Improved loss calculation" module of Fig. 4.

For each sample in a training batch, the hidden state at the final answer token position is extracted as the sample's semantic representation. The model maintains one learnable prototype vector per defect class. The contrastive loss is defined as:

$$\mathcal{L}_{con} = -\log \frac{\exp(\mathrm{sim}(h_i, p_{y_i})/\tau)}{\sum_{k=1}^{C} \exp(\mathrm{sim}(h_i, p_k)/\tau)} \tag{1}$$

where h_i is the hidden-state representation of sample i, p_{y_i} is the prototype vector for its ground-truth class, C is the total number of defect

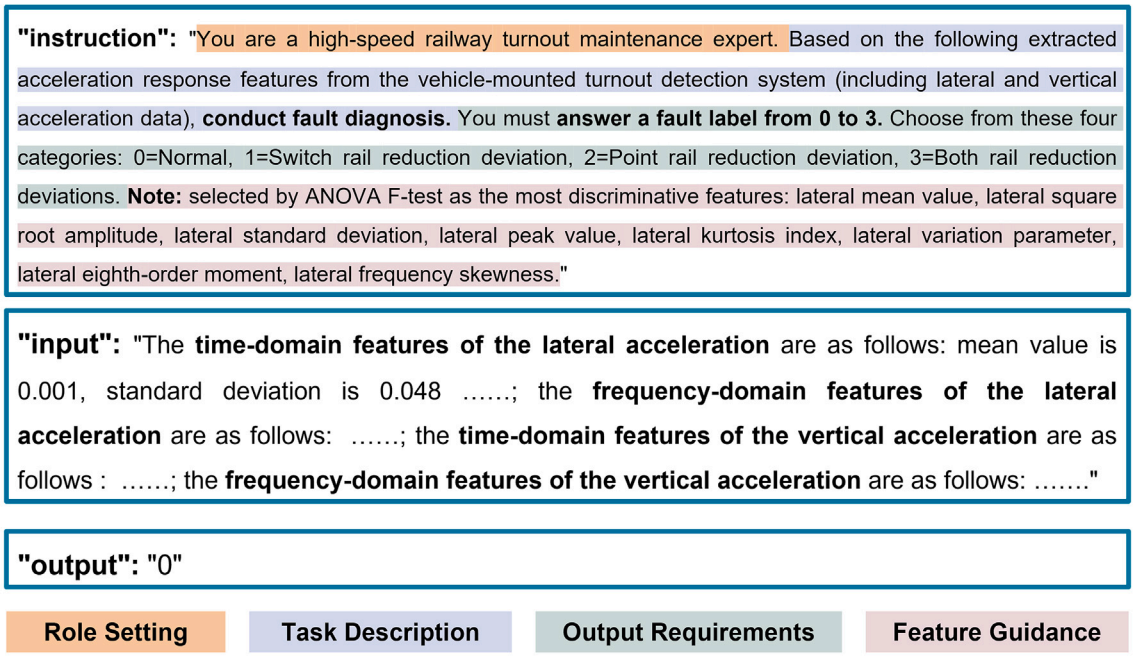


Fig. 5. Example of the text data format used for instruction fine-tuning.

categories, τ is a temperature parameter controlling the sharpness of the similarity distribution, and $\mathrm{sim}(\cdot,\cdot)$ denotes cosine similarity. Prototype vectors are updated via exponential moving average to maintain stability across training steps (Li et al., 2021):

$$p_k \leftarrow m \cdot p_k + (1-m) \cdot \bar{h}_k \tag{2}$$

where m is the momentum coefficient and $\bar{h}_k$ is the batch-mean representation for class k. The total training objective combines the standard language modelling loss with the contrastive term:

$$\mathcal{L}_{total} = \mathcal{L}_{CE} + \lambda \cdot \mathcal{L}_{con} \tag{3}$$

where λ controls the relative contribution of the contrastive term. This combined objective preserves the model's generative capabilities, which are essential for downstream report generation, while explicitly regularizing the feature space for more reliable classification.

3.1.3. Hierarchical evaluation strategy for the classification task

Because the fine-tuned LLM generates classification results as free-form text (e.g., "Based on the feature analysis, the defect category is: 2"), standard text-overlap metrics such as ROUGE-L and BLEU are poorly suited for evaluation. These metrics penalize surface variation in expression even when the categorical intent is identical—a model output of "the category is 2" and a reference of "2" would receive a low overlap score despite conveying the same diagnostic conclusion.

The study therefore adopts a hierarchical semantic parsing strategy, shown in the "Improved accuracy calculation" module of Fig. 4. After standard text preprocessing (whitespace normalization, punctuation removal), evaluation proceeds through three cascading extraction steps. The first step searches the trailing segment of the output text for a standalone integer in the valid label set {0, 1, 2, 3}; the trailing segment is prioritized because the model's final statement typically contains the conclusive answer. If no explicit integer is found, the second step matches the output against a bilingual keyword dictionary that maps natural language expressions to class indices (e.g., "level 1" → class 1, "condition zero" → class 0), covering both Chinese and English variants to accommodate cross-lingual evaluation scenarios. As a final fallback, the third step conducts a full-text scan for any digit present in the output, capturing edge cases where the model embeds the label within an unexpected sentence structure. This hierarchical approach correctly identifies semantically equivalent outputs regardless of their surface form, enabling a reliable measure of true classification accuracy. Outputs that fail all three extraction steps are recorded as prediction failures and counted as incorrect classifications.

The complete procedure, covering LoRA-based fine-tuning with the contrastive objective and hierarchical accuracy evaluation, is given in Algorithm 1.

Algorithm 1 Contrastive Instruction Fine-Tuning with Hierarchical Evaluation.

Input: Pre-trained LLM $\mathcal{M}$; training set $\mathcal{D}_{train}$; validation set $\mathcal{D}_{val}$; epochs E; learning rate η; contrastive weight λ; temperature τ; momentum m; classes C; label set $\mathcal{Y} = \{0,1,2,3\}$; keyword dictionary $\mathcal{W}$

Output: Fine-tuned adapters θ_{LoRA}; prototypes $\{p_k\}_{k=1}^{C}$

Attach LoRA adapters to $\mathcal{M}$; freeze remaining parameters; initialize $\{p_k\}$ randomly
for epoch = 1 to E **do**
  for each mini-batch $\mathcal{B} \subset \mathcal{D}_{train}$ **do**
    Forward pass: obtain logits and extract hidden state h_i at the final answer token for each sample $(x_i, y_i) \in \mathcal{B}$
    Compute $\mathcal{L}_{total} = \mathcal{L}_{CE} + \lambda \cdot \mathcal{L}_{con}$ via Eqs. (1) and (3)
    Update θ_{LoRA} via AdamW; update each p_k via Eq. (2)
  end for
  // Hierarchical validation
  for each $(x_j, y_j) \in \mathcal{D}_{val}$ **do**
    Generate $\hat{t}_j \leftarrow \mathcal{M}(x_j)$; preprocess text
    Extract $\hat{y}_j$ by: (1) trailing integer search $\in \mathcal{Y}$; (2) keyword matching via $\mathcal{W}$; (3) full-text digit scan; else mark as failure
  end for
  Compute Acc $= \frac{1}{|\mathcal{D}_{val}|} \sum_j \mathbf{1}[\hat{y}_j = y_j]$
end for
return θ_{LoRA}, $\{p_k\}_{k=1}^{C}$

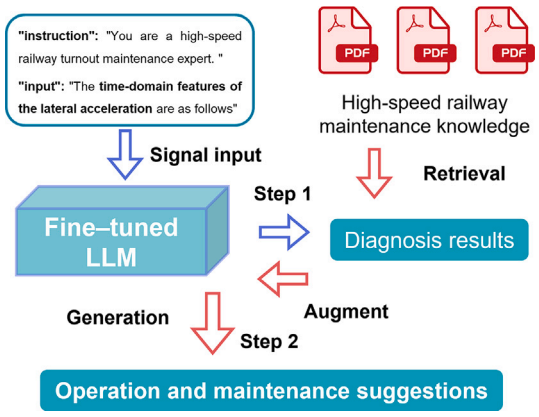


Fig. 6. Workflow of the RAG module.

Algorithm 2 RAG-Based Maintenance Recommendation Generation.

Input: Fine-tuned LLM $\mathcal{M}^*$; raw acceleration signal s; knowledge base $\mathcal{K}$; embedding model ϕ; re-ranking model ψ; number of retrieved passages K

Output: Maintenance recommendation report $\mathcal{R}$

1: *// Stage 1: Diagnostic inference*
2: Extract features from s (Table 1); construct textualized input x (Section 3.1.1)
3: $\hat{y} \leftarrow \mathcal{M}^*(x)$; $c \leftarrow \text{HierarchicalParse}(\hat{y})$ (Section 3.1.3)
4: *// Stage 2: Knowledge-grounded report generation*
5: Retrieve top-K passages: $\mathcal{P} \leftarrow \text{TopK}\big(\text{sim}(\phi(c), \phi(d)) \; \forall \; d \in \mathcal{K}\big)$
6: Re-rank: $\mathcal{P}^* \leftarrow \psi(\mathcal{P}, c)$
7: Assemble prompt $x_{gen} \leftarrow \text{Prompt}(c, \mathcal{P}^*)$; generate $\mathcal{R} \leftarrow \mathcal{M}^*(x_{gen})$
8: **return** $\mathcal{R}$

In practice, the hierarchical strategy proved highly reliable across all 26 experimental groups: the first two extraction levels (trailing integer search and keyword matching) successfully parsed every model output during both validation and testing. The third-level full-text digit scan was never triggered, owing to the explicit format constraint in the instruction prompt. Nevertheless, the theoretical risk of spurious digit matching remains for scenarios involving longer outputs or expanded label sets, and is discussed further in Section 6.

3.2. Standardized maintenance recommendation via RAG

Once the LLM has acquired defect diagnosis capability through fine-tuning, the RAG module extends its function to maintenance decision generation, closing the loop from signal acquisition to actionable output. The workflow is illustrated in Fig. 6.

The module operates in two sequential stages using the same fine-tuned model, as formalized in Algorithm 2.

Stage 1: Diagnostic inference. On-site acceleration signals are processed through the feature extraction and textualization pipeline described in Section 3.1.1, producing a structured text input. This input is fed into the fine-tuned LLM, which outputs a preliminary defect label indicating the type and severity of the reduction value deviation.

Stage 2: Knowledge-grounded report generation. The defect label from Stage 1 is used as a semantic query to retrieve the most relevant passages from a pre-built knowledge base. This knowledge base is constructed from the "Railway Engineering Technical Manual—Turnouts" through a standard document processing pipeline: the source document is segmented into passages using recursive text splitting with overlap to preserve cross-boundary context, and each passage is encoded into a dense vector representation using a dedicated embedding model. At query time, the defect label is similarly embedded, and the top-K passages with the highest cosine similarity are retrieved. A re-ranking model then reorders these candidates by contextual relevance to the specific defect type identified in Stage 1. The retrieved passages, together with the diagnostic label, are assembled into a structured prompt and passed back into the same fine-tuned LLM, which generates a maintenance recommendation that is both defect-specific and compliant with the relevant industry standards (Zhao et al., 2024).

Using a single model for both diagnostic and generative tasks reduces system complexity and avoids the coordination overhead of multimodel pipelines. Grounding the generated recommendations in an external, authoritative source ensures their professional quality and traceability, while mitigating the hallucination risk that would arise from relying solely on the model's parametric knowledge. The resulting two-stage workflow constitutes the complete signal-to-decision closed loop targeted by this study.

4. Experiment settings

This section describes the data, experimental design, and computing environment used to evaluate the proposed framework. Section 4.1 introduces the dataset and the four textualization formats. Section 4.2 details the experimental groups designed to isolate the effects of loss function, data representation, and fine-tuning hyperparameters. Section 4.3 specifies the RAG evaluation scenarios, and Section 4.4 summarizes the hardware and software environment.

4.1. Data collection and preprocessing

The dataset is derived from a stochastic multi-flexible-body dynamics model of a high-speed vehicle–turnout system developed by the research group (Tang et al., 2025). The model parametrizes uncertainties across 27 key variables in both the vehicle and turnout subsystems, and uses random irregularity spectra fitted to field measurements as the external excitation input, providing a physically representative simulation basis.

Four operating conditions are simulated: Condition 0 (nominal), Condition 1 (switch rail reduction deviation only), Condition 2 (point rail reduction deviation only), and Condition 3 (simultaneous switch and point rail deviations). For each condition, lateral and vertical acceleration time histories at key car body measurement points are recorded at a 0.25 m spatial interval throughout the full turnout passage, yielding a dataset that covers a broad range of defect types and random parameter realizations. The complete dataset contains 1500 samples in total, uniformly distributed across the four conditions; for each condition, the 375 samples are partitioned into 250 for training, 50 for validation, and 75 for testing. This balanced design simplifies the classification task relative to real-world maintenance scenarios, where defect instances are rare and class distributions are highly imbalanced; the implications of this simplification are discussed in Section 6.

Fig. 7 presents the dataset analysis in detail. The raw waveforms in panel (a) show substantial overlap between the four conditions, and the t-distributed stochastic neighbor embedding (t-SNE) projection in panel (b) confirms that the categories do not separate cleanly in the original feature space. ANOVA F-tests in panel (c) identify Lateral Root Mean Square and Lateral Standard Deviation as the two most discriminative features, providing a basis for the physical guidance prompts used in the PD and LD textualization formats. Violin plots in panel (d) further illustrate the distributional differences across conditions for these high-F features.

Four textualization formats are evaluated to assess how data representation strategies affect the model's downstream diagnostic performance. Raw data (RD) feeds the complete time series directly into the prompt, preserving all information at the cost of very high token counts and increased computational overhead. Feature data (FD) uses the extracted 24 time-frequency features, reducing input dimensionality while retaining broad coverage of the essential signal characteristics. Physical

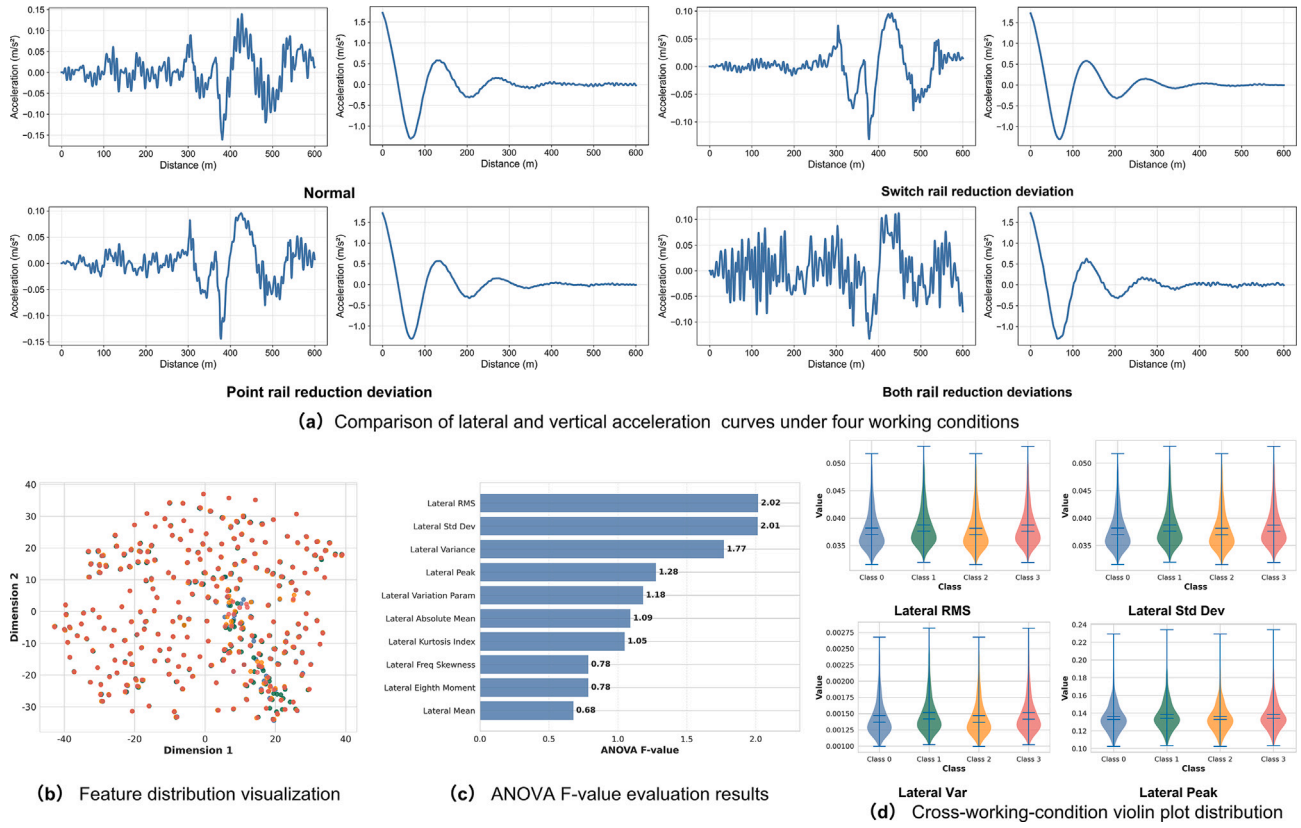


Fig. 7. Data visualization and analysis.

Table 2
Token statistics for different textualization formats.

Dataset Type	Instruction Length	Instruction Tokens	Input Length	Input Tokens	Total Length	Total Tokens
RD_ZH	134	90	40,734	38,428	40,881	38,522
RD_EN	496	96	40,797	38,430	41,306	38,530
FD_ZH	140	95	996	668	1149	768
FD_EN	550	102	1574	558	2137	665
PD_ZH	219	148	996	668	1228	821
PD_EN	827	155	1574	558	2414	718
LD_ZH	217	146	508	335	738	486
LD_EN	796	152	797	280	1606	437

data (PD) augments FD with ANOVA-guided physical prompts that direct the model's attention to the most discriminative features. Lateral data (LD) further removes the less relevant vertical acceleration features, strictly concentrating on the laterally sensitive subset identified by the ANOVA analysis to minimize potential background noise. Detailed token statistics and context length variations for each format are reported in Table 2.

4.2. Experimental setup for the diagnosis task

Experiments are organized into three groups: loss function configuration, dataset type, and fine-tuning hyperparameters. This design allows the effect of each factor to be isolated while the remaining settings are held constant. The default baseline configuration uses the LD format (lowest token count), contrastive loss weight $\lambda = 1.0$, LoRA target = 'all', learning rate = 1×10^{-4}, and effective batch size = 8. Four pre-trained base models are selected to span two architectures and two capacity levels, forming a 2×2 design matrix: Llama-3.1-8B and Llama-3.2-1B from the Llama family (Llama team, 2024), and Qwen3-8B and Qwen3-0.6B from the Qwen family (Qwen team, 2025). All fine-tuning uses LoRA (Hu et al., 2021) with rank 8, 20 training epochs, per-device batch size 2, and gradient accumulation steps 4. The full experimental design is summarized in Table 3.

4.2.1. Comparative experiments on loss functions

The first set of experiments examines whether the proposed contrastive loss outperforms standard cross-entropy training. Groups G0–G3 vary the dropout rate under pure cross-entropy loss to establish a regularization baseline. Groups G4–G7 then introduce the contrastive term at four weighting levels ($\lambda \in \{0.5, 1.0, 2.0, 3.0\}$), keeping all other hyperparameters fixed. Among these, G5 ($\lambda = 1.0$) is designated as the baseline for all subsequent experiments, as preliminary trials indicated that this weighting provides a stable balance between the language modelling and contrastive objectives. Lower λ values produced negligible representational gains, while higher values occasionally degraded validation perplexity. Hence, $\lambda = 1.0$ is adopted as the default setting for the remainder of the study.

4.2.2. Comparative experiments on dataset types

The second set of experiments assesses how data representation and language affect diagnostic accuracy. Groups G9–G11 compare FD, PD, and LD under monolingual English conditions. The RD format

Table 3
Experimental design for loss function, dataset type, and fine-tuning parameter comparisons.

Category	Group	Description	Train Data / Test Data
Loss Function	G0	CE, LD_EN	Dropout = 0.0
	G1	CE, LD_EN	Dropout = 0.1
	G2	CE, LD_EN	Dropout = 0.2
	G3	CE, LD_EN	Dropout = 0.4
	G4	CE + Contrastive, LD_EN	$\lambda = 0.5$
	G5	CE + Contrastive, LD_EN	$\lambda = 1.0$ (Baseline)
	G6	CE + Contrastive, LD_EN	$\lambda = 2.0$
	G7	CE + Contrastive, LD_EN	$\lambda = 3.0$
Dataset Type	G8	Monolingual EN	RD_EN / RD_EN (Raw Data)
	G9	Monolingual EN	FD_EN / FD_EN (Full features)
	G10	Monolingual EN	PD_EN / PD_EN (Physical guidance)
	G11 (G5)	Monolingual EN	LD_EN / LD_EN (Baseline)
	G12	ZH trained, bilingual test	FD_ZH / FD_ZH + FD_EN
	G13	ZH trained, bilingual test	PD_ZH / PD_ZH + PD_EN
	G14	ZH trained, bilingual test	LD_ZH / LD_ZH + LD_EN
	G15 (G9)	EN trained, bilingual test	FD_EN / FD_EN + FD_ZH
	G16 (G10)	EN trained, bilingual test	PD_EN / PD_EN + PD_ZH
	G17 (G5)	EN trained, bilingual test	LD_EN / LD_EN + LD_ZH
Fine-Tuning Parameters	G18 (G5)	LoRA Target, LD_EN	Target = All (Baseline)
	G19	LoRA Target, LD_EN	Target = Attention
	G20	LoRA Target, LD_EN	Target = FFN
	G21	Learning Rate, LD_EN	1×10^{-3} (High)
	G22 (G5)	Learning Rate, LD_EN	1×10^{-4} (Baseline)
	G23	Learning Rate, LD_EN	1×10^{-5} (Low)
	G24 (G5)	Batch Size, LD_EN	Effective 8 (Baseline)
	G25	Batch Size, LD_EN	Effective 25
	G26	Batch Size, LD_EN	Effective 100

(G8) was excluded because its token volume (over 38 000 tokens per sample) prevented fine-tuning completion even at batch size 1 on a single RTX 4090, confirming that raw time-series input is impractical for current-generation LLMs without prior dimensionality reduction. Groups G12–G17 evaluate cross-lingual transfer by reusing the monolingual models trained in G9–G11 (for the EN direction) and training new models on Chinese data (for the ZH direction). Each group tests the model on both language variants of the test set, enabling direct comparison between monolingual accuracy and cross-lingual transfer accuracy within a single experimental group.

4.2.3. Comparative experiments on fine-tuning parameters

The third set of experiments investigates three hyperparameters that directly govern the scope and dynamics of parameter-efficient adaptation. For the LoRA target, G18 (target = all) serves as the reference, with G19 (attention modules only) and G20 (feed-forward network (FFN) modules only) as comparisons. Learning rate variants span 1×10^{-3} (G21), 1×10^{-4} (G22, baseline), and 1×10^{-5} (G23). Effective batch sizes of 8 (G24, baseline), 25 (G25), and 100 (G26) are tested; G26 results for 8B models are unavailable due to VRAM constraints. These three factors are tested independently while all other settings remain at their baseline values.

4.3. Experimental setup for RAG

The RAG experiment uses the G5 fine-tuned model (Llama-3.1-8B) as the core inference engine, deployed locally via Ollama. The RAG pipeline is assembled in Cherry Studio, a visual orchestration interface for LLM-based workflows; Qwen3-8B handles both embedding and reranking to ensure semantic retrieval accuracy. The knowledge base is drawn from an English translation of the Chinese national standard *Railway Engineering Technical Manual—Turnouts*. Preliminary comparison indicated that Chinese and English versions of the knowledge base yield comparable report quality; the English translation was selected to maintain linguistic consistency with the base models, both of which were pre-trained predominantly on English corpora. As shown in Table 6, EN→ZH transfer consistently outperforms the reverse direction, further supporting this choice. Llama-3.2-1B and Qwen3-0.6B are excluded from the RAG evaluation because their context windows are too short to accommodate the retrieved passages alongside the diagnostic input.

Table 4
Experimental scenario design for retrieval-augmented generation.

Scenario	Workflow	Inference Model
0	Diagnosis label → direct report generation	Base model
1	Diagnosis label → RAG retrieval → generation	Base model + RAG
2	Diagnosis label → direct report generation	Fine-tuned model
3	Diagnosis label → RAG retrieval → generation	Fine-tuned model + RAG

Four scenarios are compared, as listed in Table 4. Scenarios 0 and 2 test direct report generation without retrieval, using the base and fine-tuned models respectively, while Scenarios 1 and 3 augment the same two models with RAG-based knowledge grounding. This 2×2 design separates the contributions of instruction fine-tuning and knowledge retrieval to overall report quality.

4.4. Experimental environment

All experiments were conducted in a unified cloud container comprising 13 CPU cores, 56 GB RAM, and one NVIDIA GeForce RTX 4090 (24 GB VRAM). Development and training used Visual Studio Code with dependency versions specified by the LlamaFactory framework (Zheng et al., 2024). All base model weights were sourced from Hugging Face. A single 20-epoch fine-tuning run on the LD format was completed within approximately four hours for the 8B-scale models, confirming that the proposed pipeline is feasible on commodity GPU hardware.

5. Results and analysis

Results are presented in the same order as the experimental groups defined in Section 4: loss function comparison (Section 5.1), dataset type comparison (Section 5.2), training parameter comparison (Section 5.3), RAG evaluation (Section 5.4), and generalization validation (Section 5.5).

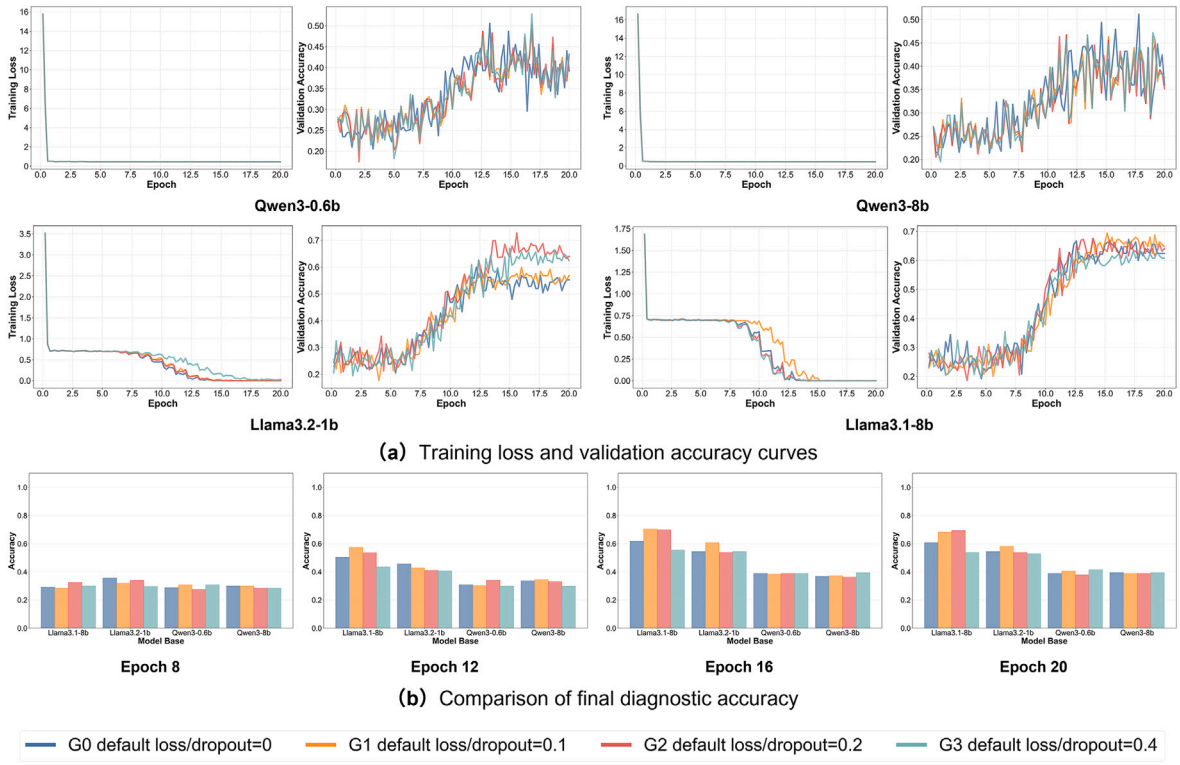


Fig. 8. Model performance with standard cross-entropy loss under different dropout rates.

5.1. Loss function comparison

Fig. 8 summarizes model behavior under pure cross-entropy loss. The training loss curves in panel (a) show rapid convergence followed by a plateau, with curves across dropout groups G0–G3 largely overlapping. Validation accuracy in panel (b) likewise shows no clear stratification across dropout rates, and final test accuracy differences within G0–G3 for any given model remain below 0.03. These results indicate that dropout adjustment alone provides no meaningful gain in this fine-tuning setting.

Introducing the contrastive loss (G4–G7, Fig. 9) produces two observable effects: a more sustained upward trend in validation accuracy across training, and a reduction in late-stage (Epoch > 15) fluctuations compared to the cross-entropy-only baseline. Both effects suggest that the contrastive term yields more stable and better-separated feature representations. Among the four weighting levels, G5 ($\lambda = 1.0$) achieves the best overall balance, with Llama-3.1-8B approaching 0.75 accuracy. Higher weights (G6, G7) do not yield further gains and in some cases slightly degrade performance, which is consistent with over-weighting the contrastive term at the expense of the language modeling objective.

5.2. Dataset type comparison

In terms of training cost, data type complexity determines computational consumption. FD and PD contain more tokens and incur higher overhead, whereas LD streamlines features to boost training efficiency. Dimensionality reduction effectively suppresses overfitting for models with limited capacity (Table 5). As noted in Section 4.2, the RD format was excluded due to excessive token length. For the remaining formats, Fig. 10 reveals a consistent pattern: the benefit of richer feature sets depends on model capacity.

For 8B-scale models, the FD format yields the highest accuracy. Llama-3.1-8B trained on FD_EN (G9/G15) reaches the study's peak test accuracy of 89.6% on the monolingual English test set (Table 6, G15 row, FD_EN column); the same model evaluated on the Chinese test set under cross-lingual transfer achieves 75.7%. This peak accuracy is a monolingual result (English training and English testing); it appears under the G15 group because G15 reuses the G9 model and additionally evaluates it on the Chinese test set to quantify cross-lingual transfer. The 13.9 percentage-point gap between the monolingual (89.6%) and cross-lingual (75.7%) evaluations of the same model directly measures the cost of language mismatch between training and inference. The greater model capacity of 8B models allows effective use of the richer FD feature set. For smaller models (Qwen3-0.6B, Llama-3.2-1B), the advantage of FD largely disappears; LD's lower redundancy enables faster convergence and comparable final accuracy despite fewer parameters. This capacity-dependent behavior can be attributed to three factors. First, small models lack the representational capacity to perform implicit feature selection when presented with 24 features, many of which are redundant for the classification task; the model cannot learn to ignore irrelevant inputs and instead fits noise in the less discriminative features. Second, the FD format produces approximately 665 tokens per sample compared with 437 for LD (Table 2), and smaller models with fewer attention heads must distribute attention across the additional tokens, diluting focus on the most discriminative quantities. Third, LD's removal of vertical features acts as an implicit regularizer, concentrating model capacity on the laterally sensitive subset that ANOVA identifies as most informative (Fig. 7(c)).

Cross-lingual transfer results (Table 6) show a consistent asymmetry: EN→ZH transfer outperforms ZH→EN transfer across all model–format combinations. For instance, Llama-3.1-8B on G15 (trained on FD_EN, cross-lingual test on FD_ZH) achieves 0.757, a 13.9 percentage-point decrease from its monolingual accuracy of 0.896 on the FD_EN test set, compared with 0.443 for the reverse direction (G12, trained on FD_ZH, tested on FD_EN). This outcome is expected given that both model families were pre-trained predominantly on English corpora. The LD format offers the most stable cross-lingual transfer, with only a 0.01 accuracy gap between English and Chinese test sets in G17, though at the cost of a lower peak accuracy.

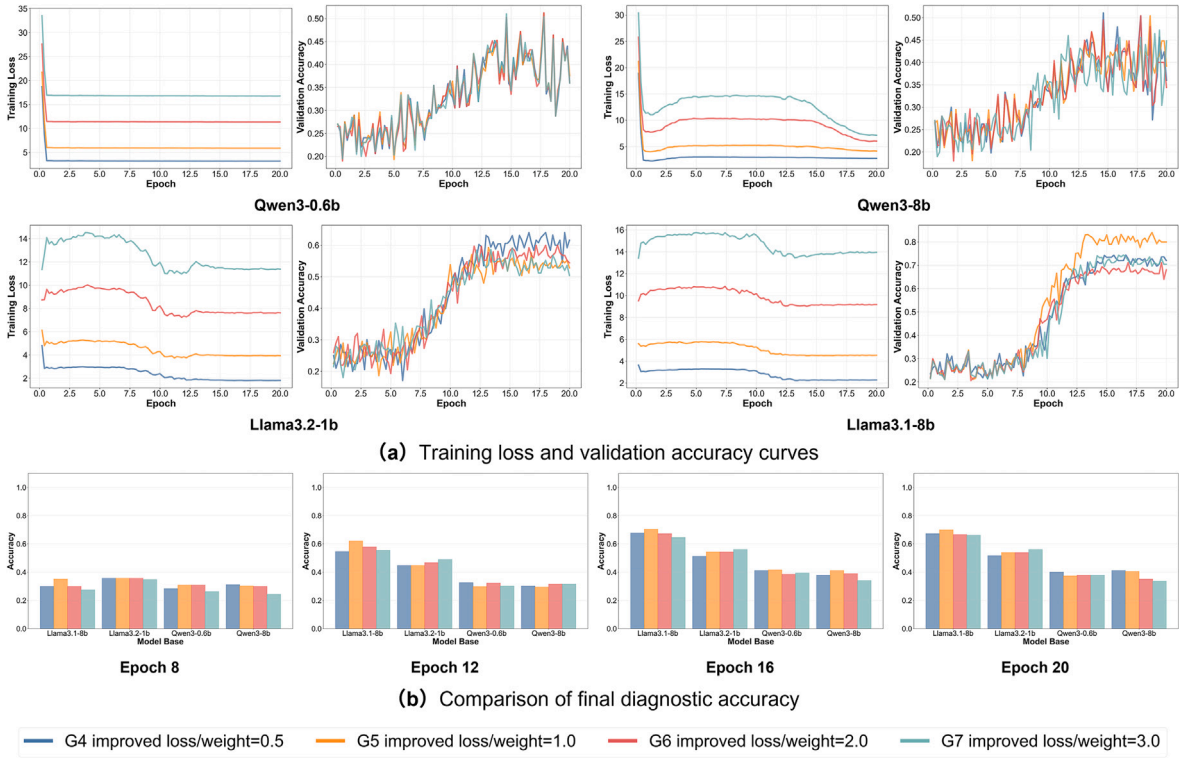


Fig. 9. Model performance with combined cross-entropy and contrastive loss under different weighting coefficients.

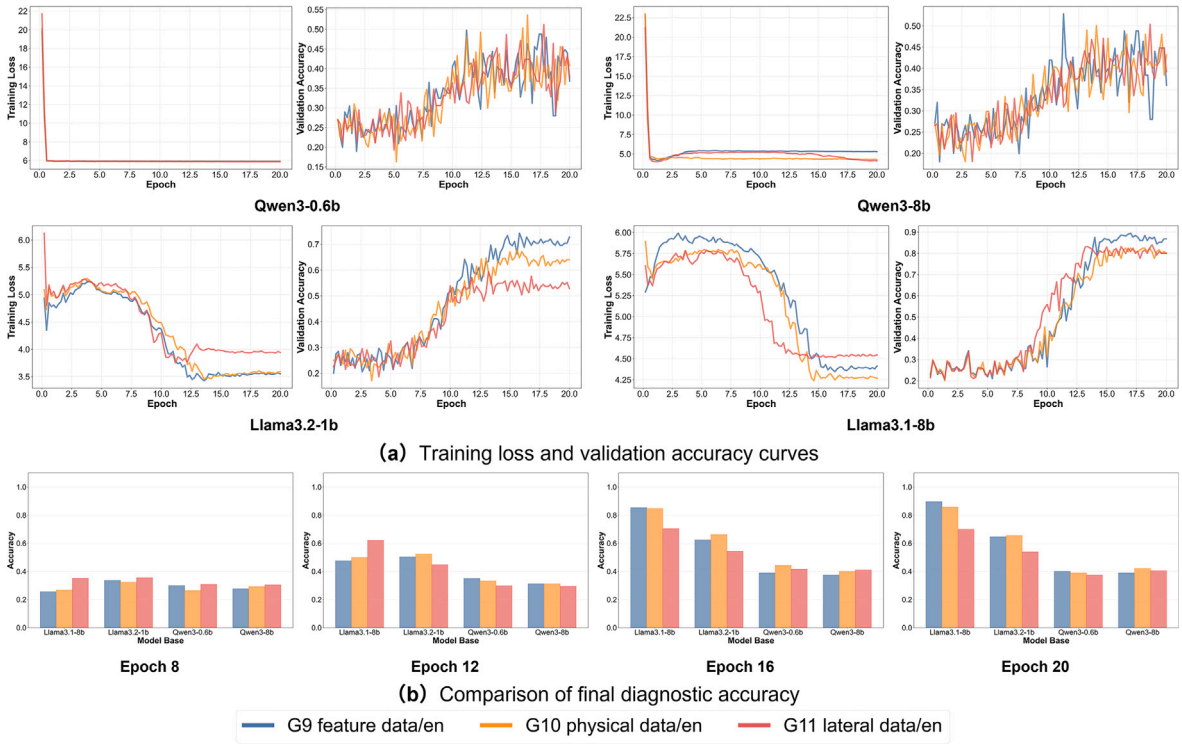


Fig. 10. Model performance with different dataset types.

5.3. Training parameter comparison

The three LoRA target variants (G18–G20) incur nearly identical computational costs (Table 7), confirming that module selection does not meaningfully affect training overhead. Performance differences, however, are appreciable: Llama-3.1-8B under G18 (target = all) achieves 0.701, compared with 0.581 for G19 (attention only) and 0.672 for G20 (FFN only). Adapting all modules simultaneously outperforms targeting either subset alone, which suggests that both attention and feed-forward layers contribute to the diagnostic mapping (Table 8).

Table 5
Training computational cost under different dataset types (FLOPs).

Training Dataset	Qwen3-0.6B	Qwen3-8B	Llama-3.2-1B	Llama-3.1-8B
G12: FD_ZH	4.15×10^{16}	7.07×10^{17}	7.99×10^{16}	6.14×10^{17}
G13: PD_ZH	4.45×10^{16}	7.58×10^{17}	8.83×10^{16}	6.79×10^{17}
G14: LD_ZH	2.65×10^{16}	4.52×10^{17}	5.54×10^{16}	4.26×10^{17}
G15: FD_EN	3.60×10^{16}	6.14×10^{17}	6.77×10^{16}	5.20×10^{17}
G16: PD_EN	3.89×10^{16}	6.63×10^{17}	7.35×10^{16}	5.65×10^{17}
G17: LD_EN	2.40×10^{16}	4.08×10^{17}	4.70×10^{16}	3.61×10^{17}

Table 6
Cross-lingual defect diagnosis accuracy.

Training	Test	Qwen3-0.6B	Qwen3-8B	Llama-3.2-1B	Llama-3.1-8B
G12: FD_ZH	FD_ZH	0.378	0.357	0.701	0.436
	FD_EN	0.384	0.416	0.603	0.443
G13: PD_ZH	PD_ZH	0.394	0.378	0.688	0.411
	PD_EN	0.389	0.368	0.640	0.432
G14: LD_ZH	LD_ZH	0.396	0.398	0.736	0.634
	LD_EN	0.357	0.395	0.608	0.571
G15: FD_EN	**FD_EN**	**0.400**	**0.389**	**0.645**	**0.896**
	FD_ZH	0.362	0.376	0.642	0.757
G16: PD_EN	PD_EN	0.389	0.421	0.656	0.858
	PD_ZH	0.320	0.394	0.603	0.704
G17: LD_EN	LD_EN	0.373	0.405	0.528	0.701
	LD_ZH	0.458	0.405	0.547	0.709

Table 7
Training computational cost under different LoRA targets (FLOPs).

Group	Qwen3-0.6B	Qwen3-8B	Llama-3.2-1B	Llama-3.1-8B	Setting
G18 (G5)	2.395×10^{16}	4.081×10^{17}	4.698×10^{16}	3.612×10^{17}	target = all
G19	2.381×10^{16}	4.073×10^{17}	4.679×10^{16}	3.606×10^{17}	target = attention
G20	2.383×10^{16}	4.077×10^{17}	4.690×10^{16}	3.609×10^{17}	target = FFN

Table 8
Test accuracy under different fine-tuning strategies.

Group	Qwen3-0.6B	Qwen3-8B	Llama-3.2-1B	Llama-3.1-8B	Setting
G18 (G5)	0.373	0.405	0.528	0.701	target = all
G19	0.378	0.389	0.517	0.581	target = attention
G20	0.373	0.373	0.539	0.672	target = FFN
G21	0.394	0.394	0.362	0.352	LR = 1×10^{-3}
G22 (G5)	0.373	0.405	0.528	0.701	LR = 1×10^{-4} (Baseline)
G23	0.325	0.378	0.352	0.373	LR = 1×10^{-5}
G24 (G5)	0.373	0.405	0.528	0.701	Batch = 8 (Baseline)
G25	0.384	0.410	0.565	0.592	Batch = 25
G26	0.389	/	0.373	/	Batch = 100

The baseline rate of 1×10^{-4} (G22) yields the best results. A tenfold increase to 1×10^{-3} (G21) causes training instability, with Llama-3.1-8B dropping to 0.352. A tenfold decrease to 1×10^{-5} (G23) leads to comparably low accuracy (Llama-3.2-1B: 0.352). The 1×10^{-4} setting therefore represents a stable optimum across the tested architectures and scales.

A moderate increase to effective batch size 25 (G25) benefits small models (Llama-3.2-1B improves from 0.528 to 0.565) through improved gradient diversity. For Llama-3.1-8B, the same setting reduces accuracy from 0.701 to 0.592, consistent with the known sensitivity of large transformers to batch size. Effective batch size 100 (G26) exceeds VRAM capacity for 8B models and also degrades small-model performance, which suggests diminishing returns from further batch size increases.

5.4. RAG results and analysis

Report quality improves progressively from Scenario 0 to Scenario 3, as shown in Fig. 11. Scenario 0 (base model, no RAG) produces outputs with vague tool references ("precision tools" without specification), absent deviation thresholds, and overly general recommendations ("regular maintenance") that lack operational specificity. Scenario 1 (base model + RAG) introduces cause analysis content and concrete operational guidance drawn from the knowledge base, substantially improving specificity. Scenarios 2 and 3 (fine-tuned model) confirm that instruction fine-tuning (applied only to fault classification, not to generation) does not degrade report generation quality: both produce logically coherent, structurally complete outputs that meet baseline expectations for a professional diagnostic report. The Scenario 3 combination of fine-tuning and RAG yields the highest quality outputs, with standards-compliant content, explicit deviation thresholds, and procedure-level recommendations that Scenario 2 cannot independently produce.

As noted in Section 4.3, the 1B- and 0.6B-scale models lack the context capacity required for the retrieval-augmented pipeline, limiting RAG-based reporting to models at the 8B scale or above.

To complement the qualitative comparison, three domain experts from universities and research institutes (none of whom are authors of this paper) independently rated the generated reports on a five-point Likert scale across five dimensions, following established practices for human evaluation of generated text (Celikyilmaz et al., 2020). Table 9 reports the averaged scores. The results indicate that RAG is the primary driver of report quality: Scenarios 0 and 2 (without RAG) score below 2.0 on all dimensions regardless of whether the model has been fine-tuned, while Scenarios 1 and 3 (with RAG) score 4.0 or above. This pattern is expected because the instruction fine-tuning in this study targets the classification task and does not directly modify the model's report-generation behavior; the quality improvement stems from the retrieval of domain-specific knowledge rather than from the fine-tuning itself.

5.5. Generalization validation

To assess the adaptability of the proposed framework beyond reduction value deviations, a preliminary generalization experiment was conducted using an independent dataset from a prior study on ballastless track arching defects (Tang et al., 2024). Unlike the reduction deviation task, which uses train body acceleration signals, the arching dataset is based on longitudinal level irregularity data collected by a track inspection system. Arching severity is determined by the difference in irregularity amplitude between a low-temperature month (January) and a high-temperature month (July): significant differences indicate arching, while minimal differences indicate normal conditions. Each sample consists of 50 measurement points at 0.25 m intervals, covering one 12.5 m track slab. The dataset contains 1800 samples across four conditions: Normal (1,000), Level 1 slight arching (400), Level 2 moderate arching (200), and Level 3 severe arching (200). The class distribution is imbalanced, reflecting a more realistic maintenance scenario than the uniformly balanced reduction deviation dataset.

Three textualization formats were evaluated. RD feeds the 50-point difference sequence directly. FD extracts 12 statistical features (mean, standard deviation, root mean square, absolute mean, peak value, skewness, kurtosis, variance, kurtosis index, peak index, waveform index, and pulse index) from the spatial difference sequence. PD augments FD with brief physical guidance on the relationship between irregularity amplitude differences and arching severity. An example of the instruction prompt and data format used for this generalization task is shown in Fig. 12. Two base models (Llama-3.1-8B and Llama-3.2-1B) were fine-tuned using the same contrastive instruction fine-tuning configuration as the best-performing reduction deviation experiments ($\lambda = 1.0$, LoRA target = all, learning rate = 1×10^{-4}, effective batch size = 8, 20 epochs).

Table 10 reports the results. The RD format yields near-chance accuracy for both models, confirming that raw irregularity sequences are too

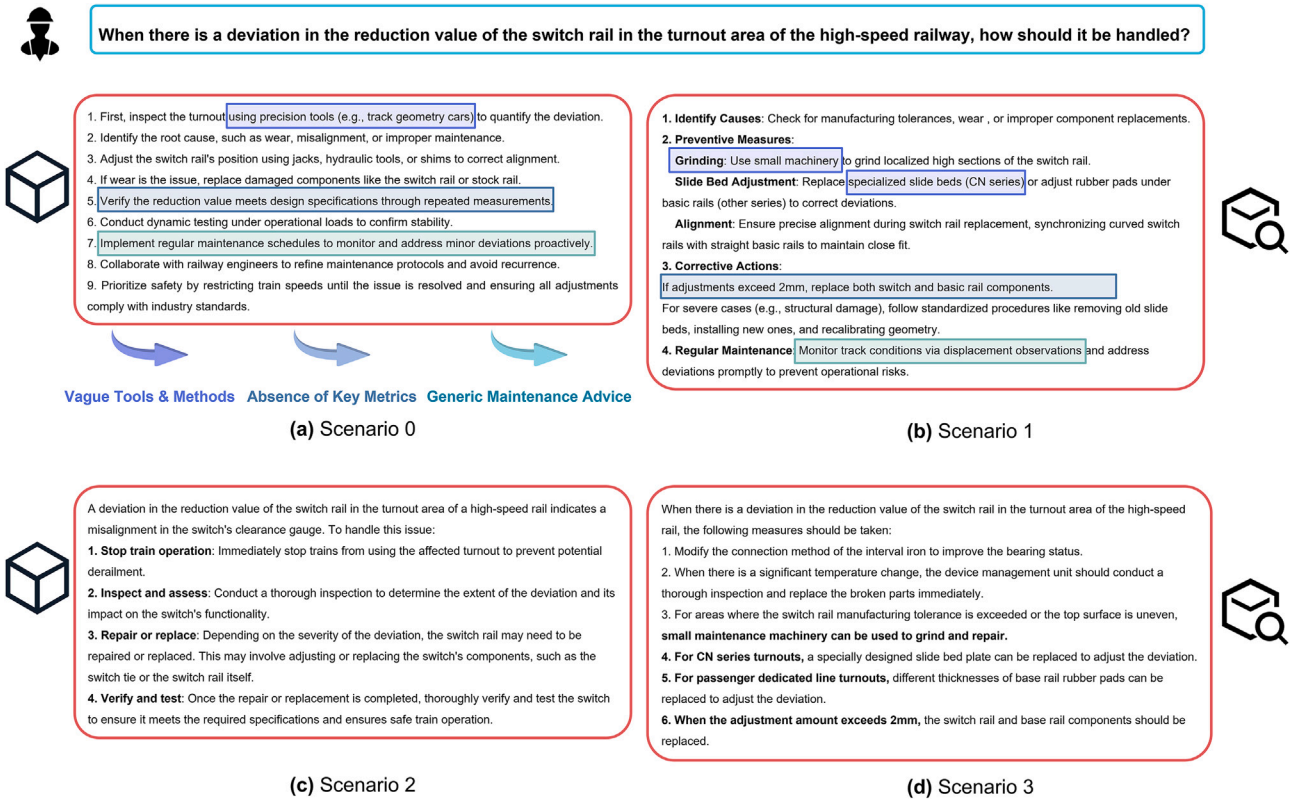


Fig. 11. Text generation results under different retrieval-augmented generation scenarios.

Table 9
Expert evaluation of generated maintenance reports (1–5 Likert scale, averaged over three experts).

Dimension	Sc. 0	Sc. 1	Sc. 2	Sc. 3
Technical Accuracy	2	4	2	4
Standards Compliance	1	4	1	4
Operational Feasibility	2	5	2	5
Completeness	1	4	1	4
Overall Quality	2	4	2	5

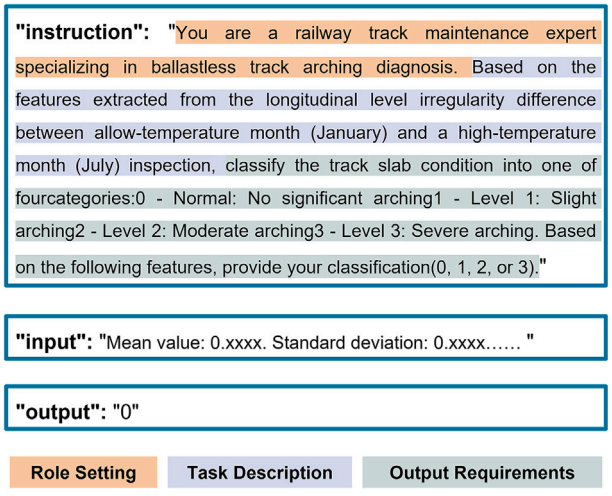


Fig. 12. Example of the text data format used for instruction fine-tuning on the arching defect dataset.

Table 10
Generalization results on the arching defect dataset.

Model	RD	FD	PD
Llama-3.1-8B	0.231	0.685	0.671
Llama-3.2-1B	0.245	0.595	0.597

long and unstructured for direct LLM processing without prior feature extraction, consistent with the RD exclusion observed in the reduction deviation experiments. The FD and PD formats achieve substantially higher accuracy, with Llama-3.1-8B reaching 68.5% on FD. The absence of an improvement from PD over FD is expected, as the physical guidance was not specifically optimized for the arching task. Notably, the class-imbalanced distribution (1000/400/200/200) in this dataset presents a harder classification problem than the balanced reduction deviation setting, yet the framework still produces meaningful diagnostic accuracy, providing initial evidence that the proposed methodology transfers across defect types and data modalities.

6. Discussion

This section discusses the main limitations of the current study and identifies directions for future research.

The dataset used in this study comprises 1500 samples generated entirely by stochastic multi-flexible-body dynamics simulation. Although the simulation model accounts for 27 parameter uncertainties and uses field-measured irregularity spectra as excitation input (Tang et al., 2025), the trained system has not been validated on real operational signals, which may exhibit different noise characteristics, sensor drift, and environmental variability. The uniformly balanced four-class distri-

bution is considerably simpler than real-world maintenance scenarios with highly imbalanced class frequencies, and the sample size, while sufficient for parameter-efficient LoRA fine-tuning, is relatively limited compared with typical LLM pretraining corpora. The most critical next step is therefore validating the framework on field-collected acceleration data from in-service HSR turnouts. When sufficient labelled field samples are available, transfer learning offers a natural adaptation pathway; when only a handful of annotated examples can be obtained under operational constraints, few-shot adaptation provides a more practical alternative. Collaboration with railway operators to access operational data under appropriate confidentiality agreements is being actively pursued.

Although the diagnostic and report-generation stages operate automatically, the overall pipeline incorporates several manually designed components: the 24 time–frequency features are predefined based on domain expertise, the instruction prompts include ANOVA-guided physical hints, the knowledge base is translated from a specific technical manual, and the hierarchical label extraction strategy relies on rule-based parsing. These choices inject domain knowledge that improves accuracy and interpretability, but they mean the framework is more accurately described as a semi-automated integrated pipeline than a fully end-to-end system. A promising direction is to integrate a convolutional or transformer-based encoder as a front-end module that projects raw acceleration signals directly into the LLM's embedding space, enabling automatic feature discovery and removing the need for manual feature selection. Separately, the third-level fallback in the hierarchical evaluation strategy (full-text digit scan, Algorithm 1) was never triggered in practice but carries a theoretical risk of matching irrelevant digits in verbose outputs; constrained decoding or structured output schemas would provide stronger guarantees for production deployment.

For the four defect categories evaluated here, a template-based approach mapping each label to a pre-written report would be a simpler alternative, and this trade-off is acknowledged. The LLM-based design is motivated by considerations that extend beyond the current four-class setting: real-world turnout maintenance involves dozens of defect types and severity levels whose combinatorial space makes templates impractical at scale; an LLM can reason over instance-specific feature values to differentiate recommendations within the same category; and the RAG mechanism allows standards updates by refreshing the knowledge base without modifying the generation logic. The present four-class evaluation thus serves as a proof-of-concept validation platform. The preliminary generalization results in Section 5.5 provide initial evidence that the framework transfers to other defect types; extending it to a comprehensive taxonomy (rail wear, contact fatigue cracking, fastener loosening, etc.) requires adapting feature sets, instruction prompts, and the RAG knowledge base, while the core infrastructure remains unchanged.

The current RAG assessment relies on qualitative comparison supplemented by expert scoring (Table 9). A comprehensive evaluation would benefit from larger-scale annotation, automated faithfulness metrics (Celikyilmaz et al., 2020), and comparison against expert-written reference reports; such evaluation is left for future work as the community establishes standardized benchmarks for domain-specific RAG. Additionally, the dense-passage retrieval currently used can exceed the context capacity of smaller models; constructing a structured Knowledge Graph from the manual's hierarchical content would enable graph-traversal-based retrieval (Graph RAG) that delivers targeted knowledge fragments with fewer tokens, potentially extending RAG to smaller-scale models. Extending the input modality to include rail surface images, acoustic emission signals, and switch machine current curves would enable multi-modal diagnosis, allowing the system to cross-validate findings across heterogeneous sensing channels. For practical deployment along railway lines, model compression is a prerequisite: reducing memory and inference latency with minimal accuracy loss, whereas knowledge distillation can transfer the diagnostic capability of the fine-tuned LLM into a compact student model better suited to edge hardware constraints.

7. Conclusion

This study proposed and validated an LLM-driven expert system for HSR turnout defect diagnosis and maintenance decision generation, addressing the persistent gap between automated data collection and actionable O&M outputs. The main findings are as follows.

(1) Domain adaptation through contrastive instruction fine-tuning. Combining a structured feature textualization strategy with a contrastive learning objective enables a general-purpose LLM to reach 89.6% diagnostic accuracy on reduction deviation defects. The contrastive term leads to more stable convergence and clearer category separation in the representation space than pure cross-entropy training with dropout regularization, while the hierarchical evaluation strategy ensures that classification intent is reliably extracted from free-form outputs.
(2) Data representation and model capacity interact. The benefit of richer feature representations (FD over LD) depends on model scale: 8B-parameter models achieve substantially higher accuracy with the full 24-feature set, whereas 1B-scale models perform comparably across formats and converge more reliably with the reduced LD representation, a finding attributable to the limited capacity of smaller models to perform implicit feature selection over redundant inputs.
(3) Integrated maintenance workflow via RAG integration. Coupling the fine-tuned diagnostic model with a domain knowledge base through RAG enables the system to produce standards-compliant, procedure-level maintenance recommendations once the diagnostic features are provided. A single-model architecture for both diagnosis and report generation avoids multi-model coordination overhead, and grounding the output in an authoritative source ensures traceability while reducing hallucination risk.

CRediT authorship contribution statement

Yi Wang: Writing – original draft, Visualization, Methodology, Conceptualization. **Xiaopei Cai:** Writing – review & editing, Resources, Project administration, Funding acquisition. **Bin Cui:** Investigation, Formal analysis. **Xueyang Tang:** Software, Data curation. **Clare Wood:** Writing – review & editing. **Yue Hou:** Writing – review & editing, Supervision.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used Claude Opus 4.6 in order to perform language polishing. After using this service, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Declaration of competing interest

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

Acknowledgements

The work was supported by the State Key Laboratory of Advanced Rail Autonomous Operation (RAO2025ZT002), Beijing Jiaotong University; Tianjin Key R&D Programme for Beijing–Tianjin–Hebei Collaborative Innovation (25YFXTHZ00260); the Natural Science Foundation of Beijing, China (L251029); the Fundamental Research Funds for the Central Universities (2025QYBS007); and the Key Research Project of China Railway Design Corporation

(2024A0253805). The authors would like to thank Associate Professor Ning Chen from Beijing University of Technology for his support of this study.

Data availability

Data will be made available on request.

References

Adeli, H. (2001). Neural networks in civil engineering: 1989–2000. *Computer-aided Civil and Infrastructure Engineering*, *16*, 126–142. https://doi.org/10.1111/0885-9507.00219

Adeli, H., & Yeh, C. (1989). Perceptron learning in engineering design. *Computer-aided Civil and Infrastructure Engineering*, *4*, 247–256. https://doi.org/10.1111/j.1467-8667.1989.tb00026.x

Aela, P., Chi, H. L., Fares, A., Zayed, T., & Kim, M. (2024). UAV-based studies in railway infrastructure monitoring. *Automation in Construction*, *167*, 105714. https://doi.org/10.1016/j.autcon.2024.105714

Amezquita-Sanchez, J. P., & Adeli, H. (2016). Signal processing techniques for vibration-based health monitoring of smart structures. *Archives of Computational Methods in Engineering*, *23*, 1–15. https://doi.org/10.1007/s11831-014-9135-7

Areerob, K., Nguyen, V. Q., Li, X., Inadomi, S., Shimada, T., Kanasaki, H., Wang, Z., Suganuma, M., Nagatani, K., Chun, P. J., & Okatani, T. (2025). Multimodal artificial intelligence approaches using large language models for expert-level landslide image analysis. *Computer-aided Civil and Infrastructure Engineering*, *40*, 2900–2921. https://doi.org/10.1111/mice.13482

Avci, O., Abdeljaber, O., Kiranyaz, S., Hussein, M., Gabbouj, M., & Inman, D. J. (2021). A review of vibration-based damage detection in civil structures: From traditional methods to machine learning and deep learning applications. *Mechanical Systems and Signal Processing*, *147*, 107077. https://doi.org/10.1016/j.ymssp.2020.107077

Cai, X., Tang, X., Pan, S., Wang, Y., Yan, H., Ren, Y., Chen, N., & Hou, Y. (2024). Intelligent recognition of defects in high-speed railway slab track with limited dataset. *Computer-aided Civil and Infrastructure Engineering*, *39*, 911–928. https://doi.org/10.1111/mice.13109

Cao, X., Xie, W., Ahmed, S. M., & Li, C. R. (2020). Defect detection method for rail surface based on line-structured light. *Measurement*, *159*, 107771. https://doi.org/10.1016/j.measurement.2020.107771

Celikyilmaz, A., Clark, E., & Gao, J. (2020). *Evaluation of text generation: A survey*. arXiv:2006.14799.

Chang, W., Cai, X., Wang, P., Wang, Q., & Sun, J. (2022). Optimizing reduced values of switch rails during the service time of high-speed railway turnouts. *Journal of Transportation Engineering, Part A: Systems*, *148*, 04022031. https://doi.org/10.1061/JTEPBS.0000689

Chen, C., Xu, T., Wang, G., & Li, B. (2020). Railway turnout system RUL prediction based on feature fusion and genetic programming. *Measurement*, *151*, 107162. https://doi.org/10.1016/j.measurement.2019.107162

Chen, J., Chen, R., Wang, P., Xu, J., An, B., Yang, F., Sun, J., & Wang, P. (2024). Wheel-rail impact and vibration characteristic frequencies at high-speed railway turnouts. *Mechanical Systems and Signal Processing*, *218*, 111537. https://doi.org/10.1016/j.ymssp.2024.111537

Deng, L., An, B., Lu, Z., Sun, Y., Wang, P., & Gao, M. (2022). Wireless monitoring of ballastless track slab deformation for high-speed railway. *IEEE Transactions on Instrumentation and Measurement*, *71*, 1–10. https://doi.org/10.1109/TIM.2022.3214268

Entezami, A., Sarmadi, H., Behkamal, B., & Mariani, S. (2025). Early warning of structural damage via manifold learning-aided data clustering and non-parametric probabilistic anomaly detection. *Mechanical Systems and Signal Processing*, *224*, 111984. https://doi.org/10.1016/j.ymssp.2024.111984

Gao, Y., & Mosalam, K. M. (2018). Deep transfer learning for image-based structural damage recognition. *Computer-aided Civil and Infrastructure Engineering*, *33*, 748–768. https://doi.org/10.1111/mice.12363

Gao, Y., Zhou, G., & Mosalam, K. M. (2026). *A Large Language Model for Disaster Structural Reconnaissance Summarization*. arXiv:2602.11588.

Hadsell, R., Chopra, S., & LeCun, Y. (2006). Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)* (pp. 1735–1742). https://doi.org/10.1109/CVPR.2006.100

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). *LoRA: Low-Rank Adaptation Of Large Language Models*. arXiv:2106.09685.

Jia, M., Cheng, Q., Tao, C., Hu, Y., Hong, Q., Cheng, W., & Liu, Z. (2025). A physics-informed train on synthetic and test on real method for evaluating large language model-generated safety-critical traffic scenarios. *Computer-aided Civil and Infrastructure Engineering*, *40*, 5153–5169. https://doi.org/10.1111/mice.70071

Jiang, Y., Wang, J., Shen, X., & Dai, K. (2025). Large language model for post-earthquake structural damage assessment of buildings. *Computer-aided Civil and Infrastructure Engineering*, *40*, 6324–6342. https://doi.org/10.1111/mice.70010

Jose, S., Nguyen, K. T. P., Medjaher, K., Zemouri, R., Lévesque, M., & Tahan, A. (2024). Advancing multimodal diagnostics: Integrating industrial textual data and domain knowledge with large language models. *Expert Systems with Applications*, *255*, 124603. https://doi.org/10.1016/j.eswa.2024.124603

La Paglia, I., Carnevale, M., Corradi, R., Di Gialleonardo, E., Facchinetti, A., & Lisi, S. (2023). Condition monitoring of vertical track alignment by bogie acceleration measurements on commercial high-speed vehicles. *Mechanical Systems and Signal Processing*, *186*, 109869. https://doi.org/10.1016/j.ymssp.2022.109869

Lee, J., Ahn, S., Kim, D., & Kim, D. (2024). Performance comparison of retrieval-augmented generation and fine-tuned large language models for construction safety management knowledge retrieval. *Automation in Construction*, *168*, 105846. https://doi.org/10.1016/j.autcon.2024.105846

Li, C., Sun, H., Li, W., Wang, Y., Wan, Z., Wu, W., Wang, P., & He, Q. (2023). A multitask learning method for rail corrugation detection using in-vehicle responses and noise data. *IEEE Transactions on Intelligent Transportation Systems*, *25*, 5054–5058. https://doi.org/10.1109/TITS.2023.3334290

Li, F., Li, X., Wen, S., & Bao, J. (2025). Multi-modal causal hypergraph reasoning for enhancing collaborative diagnosis of equipment composite failures. *Advanced Engineering Informatics*, *68*, 103611. https://doi.org/10.1016/j.aei.2025.103611

Li, J., Zhou, P., Xiong, C., & Hoi, S. C. (2021). Prototypical contrastive learning of unsupervised representations. In *Proceedings of the 9th international conference on learning representations (ICLR)*. https://doi.org/10.48550/arXiv.2005.04966

Li, Y., Trinh, H., Haas, N., Otto, C., & Pankanti, S. (2014). Rail component detection, optimization, and assessment for automatic rail track inspection. *IEEE Transactions on Intelligent Transportation Systems*, *15*, 760–770. https://doi.org/10.1109/TITS.2013.2287155

Lin, L., Zhang, S., Fu, S., & Liu, Y. (2025). FD-LLM: Large language model for fault diagnosis of complex equipment. *Advanced Engineering Informatics*, *65*, 103208. https://doi.org/10.1016/j.aei.2025.103208

Liu, Q., Li, F., Ng, K. K. H., Han, J., & Feng, S. (2025). Accident investigation via LLMs reasoning: HFACS-guided chain-of-thoughts enhance general aviation safety. *Expert Systems with Applications*, *269*, 126422. https://doi.org/10.1016/j.eswa.2025.126422

Llama team. (2024). *The Llama 3 Herd of Models*. arXiv:2407.21783.

Luo, Y., Tang, Y., Shen, C., Zhou, Z., & Dong, B. (2023). Prompt engineering through the lens of optimal control. *Journal of Machine Learning*, *2*, 241–258. https://doi.org/10.4208/jml.231023

Ma, Q., Li, T., Wang, K., Xiang, K., Chen, J., Xu, J., Chen, R., & Wang, P. (2025). Study on wheel-rail system response characteristics of rail height reduction deviations in high-speed railway turnout. *Engineering Failure Analysis*, *180*, 109876. https://doi.org/10.1016/j.engfailanal.2025.109876

Maharjan, S., & Chun, P. J. (2026). A large language model-driven framework for automated bridge specification generation and simulation validation. *Computer-aided Civil and Infrastructure Engineering*, 100014. https://doi.org/10.1016/j.cacaie.2026.100014

Qwen team. (2025). *Qwen3 Technical Report*. arXiv:2505.09388.

Sarmadi, H., Entezami, A., Yuen, K. V., & Behkamal, B. (2023). Review on smartphone sensing technology for structural health monitoring. *Measurement*, *223*, 113716. https://doi.org/10.1016/j.measurement.2023.113716

Smarsly, K., Chmelnizkij, A., Dragos, K., Peralta, J., Al-Zuriqat, T., Ahmad, M. E., Chillon Geck, C., Peralta, P., & Seitz, L. (2025). Decision making in structural health monitoring using large language models. In *Proceedings of the 15th international workshop on structural health monitoring (IWSHM 2025)*. Stanford, CA. https://doi.org/10.12783/shm2025/37344

Tang, X., Cai, X., Wang, Y., Wang, P., & Yang, F. (2025). Advanced VTSDREF for vehicle-turnout system dynamic reliability analysis: Integration of hybrid deep learning and adaptive probability density evolution method. *Reliability Engineering & System Safety*, *256*, 110762. https://doi.org/10.1016/j.ress.2024.110762

Tang, X., Wang, Y., Cai, X., Yang, F., & Hou, Y. (2024). Diagnosis of high-speed railway ballastless track arching based on unsupervised learning framework. *Computer-aided Civil and Infrastructure Engineering*, *40*, 1445–1465. https://doi.org/10.1111/mice.13342

Tao, L., Liu, H., Ning, G., Cao, W., Huang, B., & Lu, C. (2025). LLM-based framework for bearing fault diagnosis. *Mechanical Systems and Signal Processing*, *224*, 112127. https://doi.org/10.1016/j.ymssp.2024.112127

Wang, J., Li, T., Yang, Y., Chen, S., & Zhai, W. (2025). DiagLLM: Multimodal reasoning with large language model for explainable bearing fault diagnosis. *Science China Information Sciences*, *68*, 160103. https://doi.org/10.1007/s11432-024-4333-7

Wang, P., Ma, Q., Liu, J., & Xu, J. (2024). Switch rail reduction value deviation's impact on wheel–rail dynamic interaction and its efficient identification method: A numerical and experimental study. *Applied Sciences*, *14*, 12047. https://doi.org/10.3390/app142412047

Wang, X., Huang, J., Tian, Y., Sun, C., Yang, L., Lou, S., Lv, C., Sun, C., & Wang, F. Y. (2024). Parallel driving with big models and foundation intelligence in cyber–physical–social spaces. *Research*, *7*, 0349. https://doi.org/10.34133/research.0349

Wang, Y., Cai, X., Tang, X., Pan, S., Wang, Y., Yan, H., Ren, Y., & Hou, Y. (2024). HSRA-net: Intelligent detection network of anomaly monitoring data in high-speed railway. *IEEE Transactions on Intelligent Transportation Systems*, *25*, 20793–20803. https://doi.org/10.1109/TITS.2024.3477752

Wu, C., Ding, W., Jin, Q., Jiang, J., Jiang, R., Xiao, Q., Liao, L., & Li, X. (2025). Retrieval augmented generation-driven information retrieval and question answering in construction management. *Advanced Engineering Informatics*, *65*, 103158. https://doi.org/10.1016/j.aei.2025.103158

Xu, S., Zhao, K., Loney, J., Li, Z., & Visentin, A. (2025). Image-based large language model approach to road pavement monitoring. *Computer-aided Civil and Infrastructure Engineering*, *40*, 4448–4464. https://doi.org/10.1111/mice.70075

Zhang, Q., Xu, C., Li, J., Sun, Y., Bao, J., & Zhang, D. (2025). LLM-TSFD: An industrial time series human-in-the-loop fault diagnosis method based on a large language model. *Expert Systems with Applications*, *264*, 125861. https://doi.org/10.1016/j.eswa.2024.125861

Zhang, X., Gao, R., Xiao, Z., Wang, K., Liu, T., Liang, M., & Zhang, J. (2026). Natural language processing and text mining in transportation: Current status, challenges, and future roadmap. *Expert Systems with Applications*, *296*, 129050. https://doi.org/10.1016/j.eswa.2025.129050

Zhao, P., Zhang, H., Yu, Q., Wang, Z., Geng, Y., Fu, F., Yang, L., Zhang, W., Jiang, J., & Cui, B. (2024). *Retrieval-Augmented Generation for AI-Generated Content: A Survey*. arXiv:2402.19473.

Zhao, W., Yang, X., Lyu, Z., Xu, C., & Guan, Z. (2025). Road of large language model: Source, challenge, and future perspectives. *Research*, *8*, 0655. https://doi.org/10.34133/research.0655

Zheng, Y., Zhang, R., Zhang, J., Ye, Y., Luo, Z., Feng, Z., & Ma, Y. (2024). *LlamaFactory: Unified Efficient Fine-Tuning Of 100+ Language Models*. arXiv:2403.13372.

Zhou, B., Li, X., Liu, T., Xu, K., Liu, W., & Bao, J. (2024). CausalKGPT: Industrial structure causal knowledge-enhanced large language model for cause analysis of quality problems in aerospace product manufacturing. *Advanced Engineering Informatics*, *59*, 102333. https://doi.org/10.1016/j.aei.2023.102333

Zhou, X., Pan, Y., Qin, J., & Chen, J. J. (2026). Large language model-enhanced graph neural network for quantile prediction of railway track settlement near deep excavations. *Advanced Engineering Informatics*, *69*, 104105. https://doi.org/10.1016/j.aei.2025.104105