2026 European Conference on Computing in Construction
Corfu, Greece
July 12-15, 2026

PRESCRIPTIVE MAINTENANCE OF SEWER SYSTEMS BASED ON LARGE LANGUAGE MODELS

Thamer Al-Zuriqat[1], Dhananjay Birmole[1], Kosmas Dragos[1], and Kay Smarsly[1]
[1]Institute of Digital and Autonomous Construction, Hamburg University of Technology,
Blohmstraße 15, 21079 Hamburg, Germany

Abstract

Sewer systems constitute urban infrastructure critical for public health, environmental protection, and flood prevention. Decision-making in sewer systems maintenance often fails to effectively and efficiently interpret heterogeneous and multimodal inspection data into actionable maintenance strategies. To improve decision-making in sewer systems maintenance, this paper introduces a prescriptive maintenance (RxM) framework, in which a pre-trained large language model (LLM) is adapted to the domain of sewer systems using adapter-based fine-tuning. As will be shown in this paper, the LLM is capable of interpreting heterogeneous, multimodal inspection data and recommending stepwise repair procedures of sewer systems, following the paradigm of RxM.

Introduction

Civil infrastructure plays an important role in shaping modern societies by providing services essential for public well-being, such as transportation, water supply, energy, and communication networks. Maintenance of civil infrastructure is vital to ensure long-lasting, healthy, and functional infrastructure. Sewage systems comprise infrastructure for collecting and transporting sewage ("sewer systems"), and facilities for treating and disposing sewage ("wastewater treatment plants") (Davis & Cornwell, 2018). Sewer systems consist of pipelines, manholes, and pumping stations, integrated into unified networks. The operational reliability of sewer systems may be compromised by degradation due to aging, increasing hydraulic loads, and overstressing, driven by urban development and climate change (Hartmann et al., 2024).

Maintenance of sewer systems is carried out using either reactive or proactive maintenance strategies (Hartmann et al., 2025). Reactive maintenance is performed after the occurrence of significant damage, emergencies, or functional disruptions in a sewer system. By contrast, proactive maintenance comprises measures implemented independently of damage, emergencies, or functional disruptions. Proactive maintenance strategies include (i) preventive maintenance, which relies on scheduled interventions independent of condition assessment, (ii) condition-based maintenance, which is triggered by condition-assessment observations, (iii) predictive maintenance, which incorporates forecasts of degradation progression to support maintenance planning (Errandonea et al., 2020), and (iv) prescriptive maintenance, which utilizes predicted system states together with technical, operational, and contextual constraints to recommend specific maintenance actions (Smarsly et al., 2025). Within proactive maintenance strategies, reliable decision-making is crucial, particularly in prescriptive maintenance, where actions depend on the interpretation of heterogeneous and multimodal inspection data. Such multimodal inspection data includes sensor data, closed-circuit television (CCTV) footage, inspection reports, and expert assessments, which must be translated into actionable stepwise repair procedures.

A variety of approaches toward decision-making in sewer system maintenance have been proposed and can be classified into (i) traditional methods and (ii) artificial-intelligence (AI)-based methods. Traditional methods comprise visual inspections (Yang et al., 2011), standardized "defect" (i.e. damage) coding schemes, and periodic condition assessment protocols (Cao et al., 2023), which have formed the foundation of early sewer system maintenance practices. Nonetheless, traditional methods are constrained in processing heterogeneous and multimodal data, exhibit subjectivity in defect classification, and lack systematic mechanisms for integrating sensor data, visual inspection data, and textual documentation into coherent maintenance actions. AI-based methods have gained increasing attention in recent years due to advances in sensing technologies, computational resources, and data availability. AI-based methods, such as convolutional neural networks (Wang et al., 2023), "you-only-look-once" (YOLO) models (Zhang et al., 2023), single-shot multi-box detectors (Shen et al., 2023), and semantic segmentation models (Wu et al., 2025) have demonstrated effectiveness in detecting defects, classifying damage, and forecasting degradation in sewer systems. However, AI-based methods remain constrained by training-data imbalance, sensitivity to noisy inspection data, limited interpretability, and insufficient integration with established maintenance standards, resulting in outputs hardly amenable to translation into consistent, stepwise repair procedures of sewer systems. Consequently, despite the importance of reliable decision-making in prescriptive maintenance, traditional decision-making approaches often fall short of effectively and efficiently interpreting heterogeneous and

multimodal inspection data into actionable stepwise repair procedures.

To improve decision-making in sewer system maintenance, this paper introduces a prescriptive maintenance (RxM) framework, in which a pre-trained large language model (LLM) is adapted to the domain of sewer systems maintenance. In particular, the LLM is adapted to the domain of sewer systems maintenance using adapter-based fine-tuning techniques, which involves training an adapter module using a dataset consisting of repair and rehabilitation standards, subsequently integrated into the pretrained LLM. Thereupon, the LLM is capable of interpreting heterogeneous, multimodal inspection data and recommend detailed, stepwise repair procedures, adhering to the RxM paradigm. Contrary to approaches limited to prompt engineering, this study performs domain-specific fine-tuning of the Mistral-7B-Instruct-v0.2 LLM using the low-rank adaptation (LoRA) process, thereby, embedding codified normative knowledge, stemming from sewer systems maintenance repair and rehabilitation standards, into the internal representations of the LLM. The remainder of this paper is structured as follows: First, the design of the RxM-LLM framework is illuminated. Then, the implementation of the of the RxM-LLM framework is presented, followed by a validation test using a dataset not used during fine-tuning the LLM. Next, the results of the validation test are presented and discussed. Finally, the summary and conclusions as well as an outlook on potential future research directions are provided.

Design of the RxM-LLM framework

The rationale behind using LLMs for RxM of sewer systems lies in the capabilities of LLMs to self-learn complex relationship within heterogeneous and multimodal inspection data, as well as procedures mentioned in repair and rehabilitation standards. In this section, the design of the RxM-LLM framework is presented, comprising two phases, (i) preparing the fine-tuning dataset, and (ii) fine-tuning the LLM. Each phase comprises several steps, flowcharted in Figure 1.

Phase 1: Preparing the fine-tuning dataset

1. In the first step, repair and rehabilitation standards related to sewer systems maintenance are selected. Standards providing guidelines associated with inspection, repair, rehabilitation, and operational procedures for sewer systems are considered. The purpose of this step is to identify codified normative documents with technical content that can be used to form the foundation of the fine-tuning dataset.
2. In the second "preprocessing" step, standards, typically available in document formats, such as portable document format (PDF) files, that are not directly interpretable by machines are converted into a machine-readable dataset $D_{mr} = \{d_1, d_2, \ldots, d_i\}$, where d_k denotes the kth standard series part (e.g. DIN EN 1986-4) with $k = 1...i$. Each standard series part d_k is decomposed into page-wise text chunks using context-aware extraction. Specifically, each page j of d_k is represented as a text chunk $c_{k,j}$, forming the set $d_k = \{c_{k,1}, c_{k,2}, \ldots, c_{k,n}\}$, where n denotes the total number of pages in d_k. Furthermore, figures and tables appearing on page j are referenced within the corresponding text chunk $c_{k,j}$, thus preserving contextual integrity and ensuring that D_{mr} is coherent and consistently formatted.

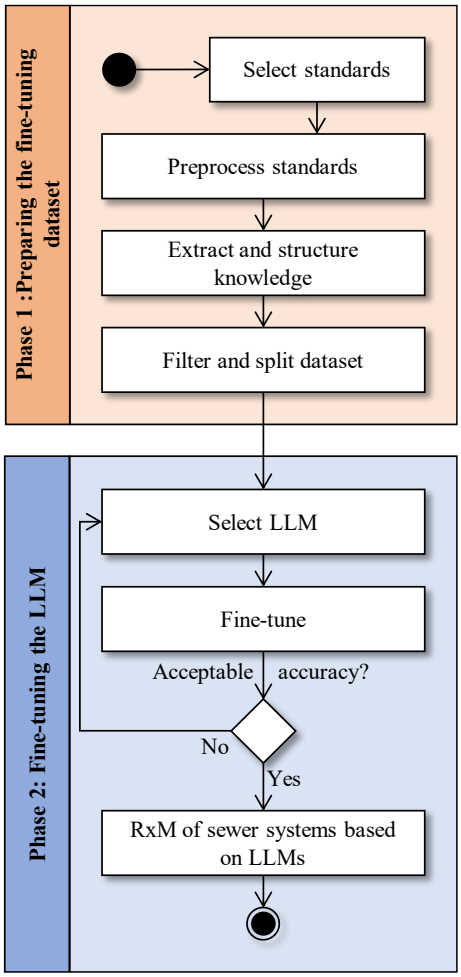

Figure 1: Flowchart of the RxM-LLM framework.

3. In the third "knowledge extraction and structuring" step, a question-answer (Q&A) dataset $D_{Q\&A}$ is generated using a two-stage strategy for fine-tuning the LLM.
   a. In the first stage, a document-level summary S_k is generated for each d_k in D_{mr}. S_k may capture the overarching scope, dominant technical topics, and normative intent of d_k. Furthermore, S_k will serve as a high-level representation of d_k and is not intended to replace the original text in d_k. Instead, S_k is used as contextual guidance to ensure that subsequent processing steps remain aligned with the overall normative context of d_k.
   b. In the second stage, each text chunk $c_{k,j}$, derived from d_k, is processed to generate Q&A pairs according to a domain-specific schema, comprising material type, damage type, damage

cause, damage location, damage severity, and potential maintenance procedures. The domain-specific schema enables a structured generation of $D_{Q\&A}$, ensuring that the Q&A pairs are context-aware and aligned with the requirements for RxM of sewer systems.

4. In the fourth step "filter and split dataset", $D_{Q\&A}$ is filtered to identify and remove similar (duplicate) entries, yielding a "filtered" Q&A dataset ($\tilde{D}_{Q\&A}$) with no redundant Q&A pairs, thereby improving the dataset diversity during fine-tuning. Next, the $\tilde{D}_{Q\&A}$ is split into a training dataset $\tilde{D}_{Q\&A(training)}$, a validation dataset $\tilde{D}_{Q\&A(validation)}$ (to avoid overfitting issues while fine-tuning the LLM), and a testing dataset $\tilde{D}_{Q\&A(testing)}$ (to eventually test the training performance with input data previously unused).

Phase 2: Fine-tuning the LLM

1. In the first step of the phase 2, candidate LLMs are selected for fine-tuning and adapting to the domain of sewer systems RxM. In general, LLMs, under the umbrella of AI systems, are trained on "big" datasets consisting of text from sources such as books, journals, and web pages, assisting the LLMs to comprehensively extract linguistic patterns, semantics, and grammar structures. Thus, LLMs are capable of analyzing, comprehending, and producing human-like language through advanced deep learning techniques, rendering LLMs well-suited for RxM applications. The candidate LLMs are assessed according to criteria, including, (i) parameter scale, relative to available computational resources, (ii) support for instruction tuning and parameter-efficient fine-tuning methods, (iii) model openness and reproducibility, and (iv) compatibility with structured Q&A generation. Then, based on the aforementioned criteria, a single LLM is selected for subsequent fine-tuning and adaptation in the next step.
2. In the second step, the LLM is fine-tuned using the $\tilde{D}_{Q\&A}$ dataset. During training, a large number of LLM parameters (sometimes in the order of billions) for predicting the next word in a sequence are adjusted, gradually learning to produce accurate, context-aware outputs. As a result, training a new LLM for sewer system maintenance would be time-consuming, and fine-tuning, i.e. retraining an already trained LLM, is employed as an alternative, essentially adapting the model to the domain of sewer systems maintenance.

3. In the final step, the accuracy of the fine-tuned LLM is checked using cross-entropy loss and mean "token" accuracy. Each token expresses the smallest unit of text processed by the LLM, obtained through sub-word tokenization, and may represent a complete word, a word fragment, a punctuation mark, or a special symbol. On the one hand, the cross-entropy loss represents the prediction error by measuring the divergence between the distribution of tokens predicted by the LLM and the ground-truth distribution, derived from the training dataset. Fine-tuning entails minimizing the cross-entropy loss between the probabilities of tokens predicted by the LLM and the reference tokens, provided by the ground truth. Low loss values indicate that the distribution of tokens predicted by the LLM closely matches the ground-truth distribution. On the other hand, the mean token accuracy serves as the principal metric for assessing fine-tuning performance, expressing the fraction of tokens predicted correctly out of all tokens in each "mini-batch" (subset of the batch, i.e. training data, processed by the LLM in a single training iteration). The mean token accuracy metric provides a granular indication of the sequence-level ability of the LLM to reproduce linguistic patterns. High mean token accuracy indicates that the LLM is capable of generating outputs closely aligned with realistic sewer systems maintenance procedures.

The combination of the two phases results in actionable information, facilitating real-time decision making in sewer system maintenance. In what follows, the implementation of the RxM-LLM framework is presented.

Implementation of the RxM-LLM framework

The implementation is presented in accordance with the phases of the RxM-LLM framework, previously described, i.e. "preparing the fine-tuning dataset" and "fine-tuning the LLM". In this context, standards on sewer systems repair and rehabilitation, issued by the German Institute for Standardization (Deutsches Institut für Normung, DIN) and harmonized as European standards (Europäische Normen, EN), as well as standards issued by the German Association for Water, Wastewater and Waste (Deutsche Vereinigung für Wasserwirtschaft, Abwasser und Abfall, DWA) are incorporated into the implementation. Furthermore, the programming language Python is employed for implementing both phases.

Phase 1: Preparing the fine-tuning dataset

Figure 2 illustrates the standard series, selected in step 1, used in this study. The parts from the standards series selected, specifically 33 parts, are acquired through the databases "DWA Regelwerk" (Deutsche Vereinigung für Wasserwirtschaft, Abwasser und Abfall (DWA), n.d.) and the standards database "Nautos" (Nautos, n.d.) as PDF files. The standards include:

- The DIN EN 13508 series, which constitute a Europe-wide-harmonized basis for governing inspection, assessment, and documentation of sewer and drainage systems (DIN EN 13508),
- The DIN EN 1986 series, which regulates the planning, design, construction, operation, and maintenance of drainage systems located on private properties, ensuring alignment between private and municipal sewer systems (DIN EN 1986),
- The DWA-M 143 series, which defines technical procedures for inspecting, assessing, and rehabilitating sewer systems (DWA-M 143),

- The DWA-M 144 series, which specifies technical requirements, procedures, and evaluation principles for structural condition assessment of sewer pipelines and manholes, providing a uniform basis for damage classification and maintenance planning (DWA-M 144), and
- The DWA-M 149 series, which provides methods and technical criteria for hydrological, hydraulic, and performance-based assessment of sewer systems, supporting reliable network analysis, monitoring, and operational planning (DWA-M 149).

Data preprocessing (step 2) for converting the PDF files into the machine-readable dataset $D_{mr} = \{d_1, d_2, \ldots, d_{33}\}$ is achieved via a Python preprocessing pipeline. Within the pipeline, textual content contained in the PDF files is extracted into a textual corpus, using optical character recognition (OCR). The textual corpus is then segmented into page-wise text chunks, such that each page j of the standard part d_k corresponds to one text chunk $c_{k,j}$. In parallel, figures contained in the PDF files are exported as separate portable network graphics (PNG) files with automatically assigned captions, and tables are reconstructed as spreadsheets. Figures and tables identified on page j are referenced within the corresponding $c_{k,j}$ to preserve contextual integrity between textual content and visual or tabular information.

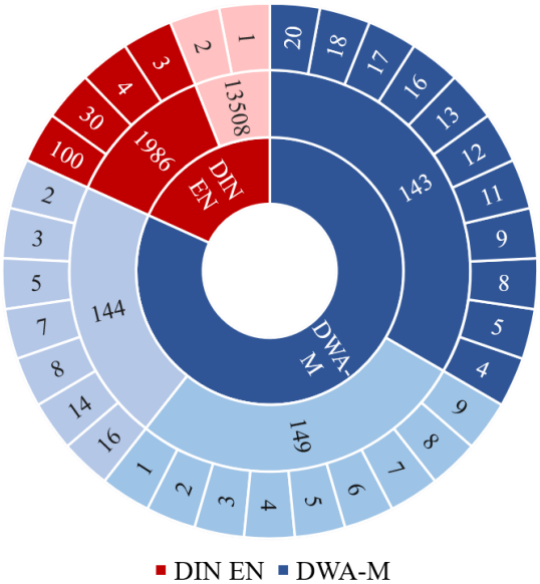

Figure 2: Standards series selected for the implementation of the RxM-LLM framework.

Knowledge extraction and structuring (step 3) for transforming D_{mr} into $D_{Q\&A}$ is accomplished as follows: First, a generation prompt based on Chat GPT-4o-mini LLM is used to produce document-level summaries $S = \{S_1, S_2, \ldots, S_{33}\}$ of D_{mr}, where the summary S_k corresponds to the kth standard part. Next, for standard part d_k, a dedicated generation prompt based on GPT-4o-mini (33 generation prompts in total) is used for generating Q&A pairs using the page-wise text chunks c_{jk}. The GPT-4o-mini generation prompts are constrained by the domain-specific schema, defined in the second stage of step 3, as well as by the document-level summaries $S = \{S_1, S_2, \ldots, S_{33}\}$. As a result, a total of 9,509 Q&A pairs are generated for $D_{Q\&A}$. Figure 3 presents a breakdown of the total Q&A pairs generated from each standard series using the GPT-4o-mini generation prompts.

In the final step of the first phase, filtering and splitting, $D_{Q\&A}$ is filtered to identify and remove duplicate Q&A pairs. Following user review and a cosine-similarity filtering, a total of 476 Q&A pairs are removed. As a result, $\tilde{D}_{Q\&A}$ comprises 9,033 Q&A pairs, which are used for fine-tuning the adapter in the next step. Finally, $\tilde{D}_{Q\&A}$ is split into 80% for training ($\tilde{D}_{Q\&A,training}$) (7,226), 15% for validation ($\tilde{D}_{Q\&A,validation}$) (1,355), and 5% for testing ($\tilde{D}_{Q\&A,testing}$) (452). From $\tilde{D}_{Q\&A,testing}$ a total of 200 Q&A pairs are held-out (not used to evaluate the fine-tuning process) and used later in the validation section. Thus, $\tilde{D}_{Q\&A,testing}$ comprises 252 Q&A pairs.

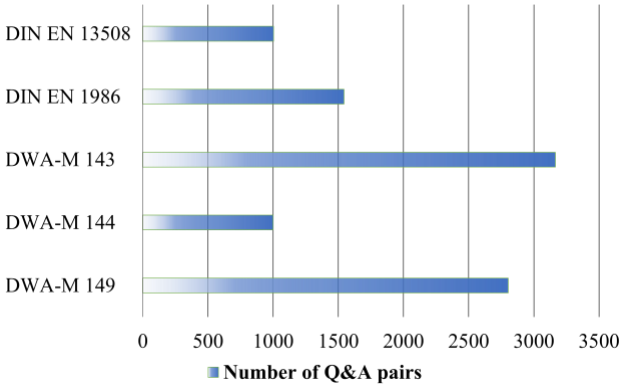

Figure 3: Number of Q&A pairs generated per standard series.

Phase 2: Fine-tuning the LLM

In the LLM selection step, which constitutes the first step of phase 2, two candidate LLMs are initially considered, namely DeepSeek-VL2 and Mistral-7B-Instruct-v0.2 because of multimodal processing and instruction-based text generation, respectively. DeepSeek-VL2 is a multimodal-vision LLM, considered due to the ability of the model to process visual inputs, such as CCTV data, figures from standards, and textual inspection reports. Mistral-7B-Instruct-v0.2 is an instruction-tuned LLM, designed for instruction following and structured text generation. Based on the criteria defined in step 1 of phase 2, the following rationale governs the selection of the LLM: While DeepSeek-VL2 provides a native multimodal processing, the parameter scale and hardware requirements needed for DeepSeek-VL2 exceed the available computational resources. By contrast, Mistral-7B-Instruct-v0.2 supports LoRA through open-source frameworks and enables efficient fine-tuning through libraries such as Hugging Face Transformers, supervised fine-tuning frameworks, and bitsandbytes, under constrained hardware conditions. As a result, the Mistral-7B-Instruct-v0.2 model is selected.

Fine-tuning (step 2 of phase 2) of the Mistral model is achieved via parameter-efficient LoRA. The Fine-tuning is conducted on a single NVIDIA RTX 4090 GPU. The baseline model, i.e. Mistral-7B-Instruct-v0.2, is loaded

using 4-bit NF4 with double quantization, while computations are carried out in bfloat16 to reduce memory consumption. Optimization is performed using the paged AdamW optimizer in 8-bit mode with a learning rate of $2 \cdot 10^{-4}$ and a linear learning-rate schedule. A warm-up phase covering 3% of the total training steps and a weight decay of 0.01 are applied. Training of the adapter is conducted using gradient accumulation to obtain an effective batch size of 32 token sequences, based on a per-device mini-batch size of four token sequences. Input sequences are packed to a maximum length of 2,048 tokens. The training process is configured for a maximum of three epochs, with gradient checkpointing enabled to reduce memory consumption. Moreover, early stop is applied once validation loss (from applying the $\tilde{D}_{Q\&A(validation)}$ dataset) fails to improve two consecutive times. As shown in Figure 4, the fine-tuning process converges to a stable state, with the validation cross-entropy loss stabilizing between 0.32 and 0.34 and the mean token accuracy reaching approximately 88%. The results of the fine-tuning indicate that a consistent and stable adaptation of the Mistral model to the domain-specific task of sewer systems maintenance.

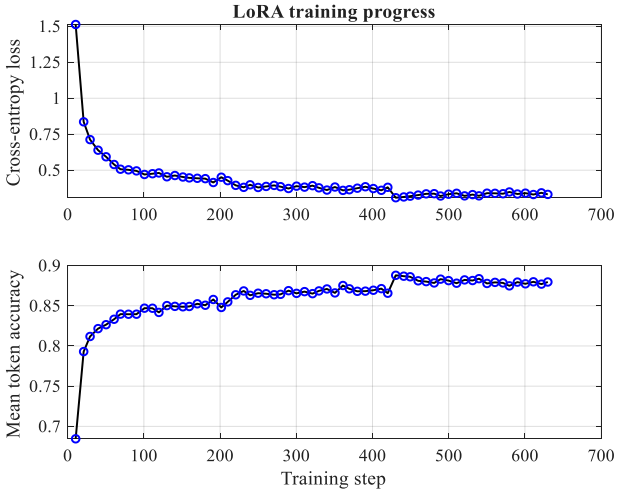

Figure 4. Performance of the LLM fine-tuning, in terms of cross-entropy loss and mean token accuracy.

In the final step, the dataset $\tilde{D}_{Q\&A(testing)}$ is used upon completing training to verify the accuracy of the fine-tuning process of the "adapted" Mistral model, based on the cross-entropy loss and the mean token accuracy, showcasing that 91.82% of the tokens have been correctly predicted. In the following section, the adapted Mistral model, i.e., the Mistral model equipped with the LoRA adapter, is validated and compared against a baseline Mistral model (Mistral-7B-Instruct-v0.2) without adapters.

Validation test

In this section, the RxM-LLM framework is validated by assessing the capabilities of the adapted Mistral model, developed in the implementation section, to interpret heterogeneous, multimodal inspection data, previously unseen, and generate detailed stepwise repair procedures of sewer systems maintenance. In the validation test, a total of 200 Q&A pairs are selected from a held-out dataset not used during fine-tuning. The Q&A pairs cover distinct damage scenarios, damage locations, and repair severity levels. Furthermore, answers generated by the adapted Mistral model are compared with those produced by a baseline Mistral model without adapters, thereby enabling an assessment of how well the adapted Mistral model fits to the domain of sewer systems maintenance.

Both the adapted Mistral model and the baseline Mistral model are evaluated using a domain-specific scoring rubric, which serves as the criterion for assessing the suitability of the models for sewer systems RxM. The scoring rubric is designed to capture requirements of sewer systems maintenance workflows and comprises the following criteria: (i) Structural clarity of responses, assessing whether answers are presented in stepwise and operationally executable formats; (ii) conformity with standards, evaluating explicit references to relevant repair and rehabilitation standards series; (iii) material and damage specificity, assessing the differentiation between pipeline materials (e.g., concrete, vitrified clay, polyvinyl chloride, and glass reinforced plastic) and common damage types (e.g., corrosion, root intrusion, and infiltration); (iv) actionability, evaluating the inclusion of repair and rehabilitation measures, such as CCTV inspection, high-pressure cleaning, injection grouting, short-liner techniques, or cured-in-place pipe lining; (v) consideration of safety and operational constraints, including confined-space entry, bypass pumping, and gas monitoring; (vi) verification and quality assurance measures, such as leak testing, deflection testing, post-rehabilitation CCTV inspection, or as-built documentation; (vii) numerical specificity, assessing the presence of quantitative information included in the repair and rehabilitation standards, such as tolerances or permissible deformation limits; and (viii) a hallucination penalty, applied to answers introducing non-domain-specific or implausible repair and rehabilitation procedures, such as irrelevant inspection technologies or incoherent procedural fragments. The scoring rubric quantifies the performance of each model on a scale from 0 to 10. The results of the validation test are presented and discussed in the following section.

Results and discussion

The quantitative performance of the both models for the 200 Q&A pairs is summarized in Table 1. The evaluation scores are computed using a LLL acting as a judge (LLM-as-a-judge), specifically GPT-5o, which assesses the generated responses based on the predefined rubric criteria introduced in the previous section. The baseline Mistral-7B-Instruct-v0.2 model exhibits an average score of 4.8/10 (±1.2), reflecting frequent fragmentation and limited actionability despite occasional references to standards. Meanwhile, the adapted Mistral model has achieved a higher overall performance with an average score of 7.6/10 (±1.0). The adapted Mistral model demonstrates a strong domain specificity and structural clarity, with correct material and damage references present in 71% of responses.

Table 1: Performance of the adapted Mistral model as well as the baseline Mistral model for the 200 Q&A pairs

Model	Score (1-10)	Standard deviation
Baseline Mistral model	4.8	±1.2
Adapted Mistral model	7.6	±1.0

In addition, more than 65% of the outputs generated by the adapted Mistral model, included at least three distinct repair and rehabilitation measures. A detailed breakdown of the performance of the two models across the individual evaluation criteria, mentioned above, is illustrated in Figure 5.

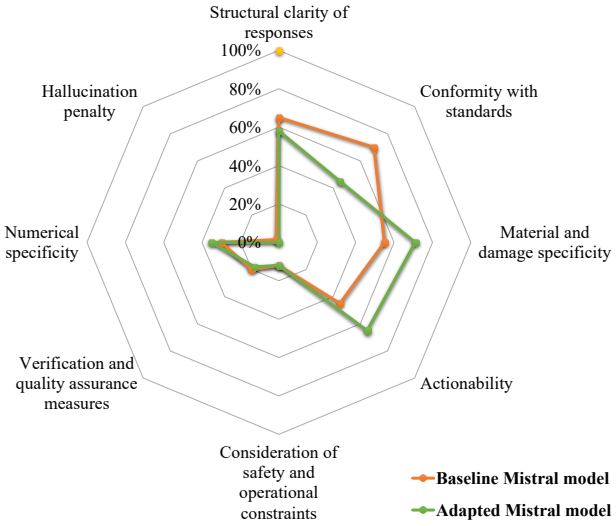

Figure 5. Performance of the two models, based on individual criteria.

Finally, to qualitatively illustrate differences between the baseline Mistral model and the adapted Mistral model, an exemplary prescriptive maintenance query is evaluated across both models. The example addresses the rehabilitation of radial cracks in vitrified clay pipes using short glass-fiber liners, representing a typical maintenance task that requires material awareness, damage-specific repair selection, and operational sequencing. The responses generated by both models are shown in Figure 6.

The baseline Mistral model demonstrates general instructional capability but lacks the technical insight required for RxM of sewer systems. The responses generated by the baseline Mistral model remain abstract and primarily describe general principles rather than actionable procedures. In addition, in outputs generated by the baseline Mistral model, damage-repair coupling is weak, and outputs frequently lack material-specific adaptation and operational details. Furthermore, stepwise repair procedures execution is incomplete, and references to safety measures and quality assurance are often omitted. Finally, the baseline Mistral model provides a superficial overview of repair and rehabilitation concepts without sufficient procedural specificity, rendering the output unsuitable for real-world RxM applications of sewer systems.

The adapted Mistral model shows a clear improvement over the baseline Mistral model, driven by damage-repair mappings introduced through structured fine-tuning. The outputs generated by the adapted Mistral model demonstrate strong alignment with technical procedures and high domain awareness. Furthermore, the adapted Mistral model generates detailed, stepwise repair and rehabilitation procedures, and consistently accounts for sewer pipes material, defect types, and conditions severity. In addition, the adapted Mistral model produces outputs resembling technical repair and rehabilitation reports rather than generic advisory text. In addition, repair and rehabilitation techniques such as cured-in-place pipe lining, short-liner installation, robotic sealing, injection grouting, packer-based repairs, and reinstatement are frequently and appropriately referenced. The adapted Mistral model exhibits accurate matching between damage mechanisms and repair methods, indicating strong alignment with sewer systems repair and rehabilitation practices.

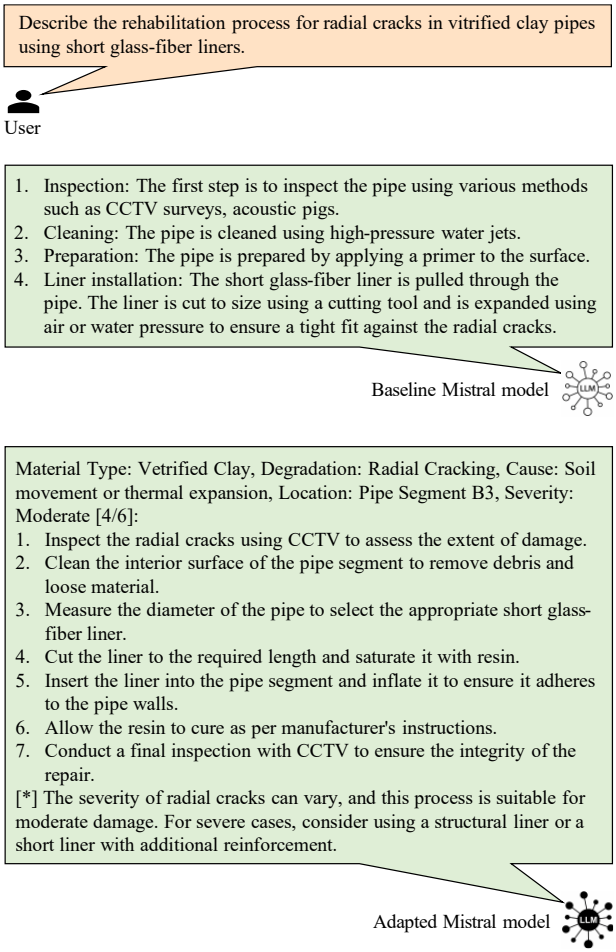

Figure 6: Example conversation of a user with the baseline Mistral model and the adapted Mistral model.

Summary and conclusions

Sewer systems constitute urban infrastructure critical for public health, environmental protection, and flood prevention. Sewer systems maintenance is vital to ensure long-lasting, healthy, and functional infrastructure.

However, decision-making in sewer systems maintenance has failed to effectively interpret heterogeneous and multimodal inspection data into actionable maintenance strategies. This paper has presented a prescriptive maintenance framework, in which, a pre-trained large language model has been adapted to the domain of sewer systems maintenance. The RxM-LLM framework comprises two phases, (i) preparing the fine-tuning dataset, and (ii) fine-tuning the LLM, and it has been implemented based on sewer systems standards defined under the DIN EN and DWA standards. In the implementation of the RxM-LLM framework, a generation prompt based on the ChatGPT-4o-mini model has been employed to generate document-level summaries for each standard. Thereupon, the summaries, together with a domain-specific schema, have been used to constrain downstream generation prompts for producing structured Q&A pairs. The dataset has served as an input for low-rank adaptation fine-tuning of the Mistral-7B-Instruct-v0.2 model. In the validation test, capabilities of the adapted Mistral model in interpreting heterogeneous, multimodal inspection data, and generate detailed stepwise repair procedures of sewer systems maintenance have been evaluated and compared to a baseline Mistral model. The validation test results has shown that the adapted Mistral model outperformed the baseline Mistral model. The adapted Mistral model has demonstrated enhanced structural clarity, consistently generating material-aware, damage-specific, and actionable repair and rehabilitation procedures. The results have demonstrated that adapter-based fine-tuned LLMs are capable of correctly interpreting heterogeneous and multimodal inspection data and recommend stepwise repair procedures, thus supporting decision-making in sewer systems maintenance. Future work may involve further developing the RxM-LLM framework to improve uncertainty handling and additional validation tests using real-world sewer systems inspection data.

Acknowledgments

This work is funded by the German Research Foundation (DFG) under grant SM 281/33-1 and by the German Federal Ministry of Transport (BMV) under grant 01FV2059C. The financial support is gratefully acknowledged. Any opinions, findings, conclusions, or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of BMDV or DFG.

References

Cao, G., Guo, S., Wei, J., Huang, R., & Li, M. (2023). Comparison of four sewer condition assessment protocols based on field data. Water, 15(21), 3790.

Davis, M.L. & Cornwell, D.A. (2013). Introduction to environmental engineering (5th ed.). McGraw-Hill Education. New York, NY, USA.

Deutsches Institut für Normung (DIN). DIN EN 13508: Investigation and assessment of drain and sewer systems outside buildings – Parts 1 and 2. Berlin, Germany: Beuth Verlag.

Deutsches Institut für Normung (DIN). DIN EN 1986: Drainage systems for buildings and land – Parts 3, 4, 30, and 100. Berlin, Germany.

Deutsche Vereinigung für Wasserwirtschaft, Abwasser und Abfall (DWA). DWA-M 143: Sanierung von Entwässerungssystemen außerhalb von Gebäuden – Teile 4, 5, 8, 9, 11, 12, 13, 16, 17, 18, und 20. Hennef, Germany.

Deutsche Vereinigung für Wasserwirtschaft, Abwasser und Abfall (DWA). DWA-M 144: Zusätzliche Technische Vertragsbedingungen (ZTV) für die Sanierung von Entwässerungssystemen außerhalb von Gebäuden – Teile 2, 3, 5, 7, 8, 14, und 16. Hennef, Germany.

Deutsche Vereinigung für Wasserwirtschaft, Abwasser und Abfall (DWA). DWA-M 149: Zustandserfassung und -beurteilung von Entwässerungssystemen außerhalb von Gebäuden – Teile 1, 2, 3, 4, 5, 6, 7, 8, und 9. Hennef, Germany.

Deutsche Vereinigung für Wasserwirtschaft, Abwasser und Abfall (DWA). DWA-Regelwerk: Technische Regeln Wasserwirtschaft, Hennef, Germany. Available at: https://de.dwa.de/de/regelwerk-fachpublikationen.html (Accessed: 30 April, 2025).

Errandonea, I., Beltrán, S., & Arrizabalaga, S. (2020). Digital Twin for maintenance: A literature review. Computers in Industry, 123(2020), 103316.

Hartmann, S., Al-Zuriqat. T., Peralta, P., Valles, R., Gölzhäuser, P., Al-Hakim, Y., Mackenbach, S., Klemt-Albert, K., & Smarsly, K. (2024). Digital twins for sewer systems maintenance: A systematic literature review. In: Proceedings of the 24th International Conference on Construction Applications of Virtual Reality (CONVR). Sydney, Australia, November 4-6, 2024, pp.345-357.

Hartmann, S., Valles, R., Schmitt, A., Al-Zuriqat, T., Dragos, K., Gölzhäuser, P., Jung, J. T., Villinger, G., Varela Rojas, D., Bergmann, M., Pullmann, T., Heimer, D., Stahl, C., Stollewerk, A., Hilgers, M., Jansen, E., Schoenebeck, B., Buchholz, O., Papadakis, I., Merkle, D. R., Jäkel, J.-I., Mackenbach, S., Klemt-Albert, K., Reiterer, A., & Smarsly, K. (2025). Digital-twin-based management of sewer systems: Research strategy for the KaSyTwin project. Water, 17(3), 299.

Nautos (n.d.) Nautos database for standards, technical regulations and guidelines. Berlin, Germany. Available at: https://nautos.de (Accessed: 30 April, 2025).

Shen, D., Liu, X., Shang, Y. and Tang, X. (2023). Deep learning-based automatic defect detection method for sewer pipelines. Sustainability, 15(12), 9164.

Smarsly, K., Chmelnizkij, A., Dragos, K., Peralta, J., Al-Zuriqat, T., Ahmad, M. E., Chillón 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). Stanford, CA, USA, September 09-11, 2025, pp.641-648.

Wang, X., Thiyagarajan, K., Kodagoda, S. and Zhang, M. (2023). PIPE-CovNet: Automatic in-pipe wastewater infrastructure surface abnormality detection using convolutional neural network. IEEE Sensors Letters, 7(4), pp.1-4.

Wu, Z., Li, M., Han, Y. and Feng, X. (2025). Semantic segmentation of 3D point cloud for sewer defect detection using an integrated global and local deep learning network. Measurement, 253(2025), 117434.

Yang, M.D., Su, T.C., Pan, N.F., &Yang, Y.F. (2011). Systematic image quality assessment for sewer inspection. Expert Systems with Applications, 38(3), pp.1766-1776.

Zhang, J., Liu, X., Zhang, X., Xi, Z. and Wang, S. (2023). Automatic detection method of sewer pipe defects using deep learning techniques. Applied Sciences, 13(7), 4589.