Improved Information Extraction from Bridge Inspection Reports using Fine-tuned Generative Pre-trained Transformers

Abdelhady Omar[1] and Osama Moselhi[2]

[1] Ph.D. Candidate, Department of Building, Civil and Environmental Engineering, Concordia University, Montréal, QC, Canada.
[2] Professor and Director, Centre for Innovation in Construction and Infrastructure Engineering and Management (CICIEM), Department of Building, Civil and Environmental Engineering, Concordia University, Montréal, QC, Canada.
Email: abdelhady.omar@mail.concordia.ca, moselhi@encs.concordia.ca

Abstract –

Bridge inspection reports contain a wealth of crucial data on bridge components and their related structural defects. This study introduces a novel method that harnesses the power of Generative Pre-trained Transformers (GPT) for improved information extraction from bridge inspection reports. While most studies in this domain focus solely on data extraction, this study transforms inspection data into a ready-to-use format for better utilization in condition assessment and predictive modeling, enabling better-informed budget allocation and decision-making. It employs (1) a baseline GPT model (BL-GPT), which leverages OpenAI's large language models to process textual inspection data through optimized prompt engineering, and (2) a fine-tuned GPT model (FT-GPT), which enhances the baseline by incorporating domain-specific training to improve performance. These models capture and evaluate the severity levels of reinforced concrete bridge defects based on textual inspection data. The models are validated on data extracted from 2,255 inspection reports—spanning a period of five years (2018–2022)—for a set of bridges in Québec, Canada. The FT-GPT is found to significantly improve performance, stability, and reliability in detecting and standardizing the severity of different types of concrete defects in bridge decks. In specific, it achieves accuracy rates of 98.79% for rebar corrosion, 99.09% for delamination, and 98.64% for cracking, scaling, and spalling of concrete. This study demonstrates the potential integration of generative AI in asset management, an application that has yet to be realized.

Keywords –

Bridge Inspection Reports; Infrastructure Asset Management; Concrete Bridge Decks; Natural Language Processing (NLP); Large Language Models (LLMs); Generative AI (GenAI); Generative Pre-trained Transformers (GPTs); Prompt Engineering; Fine-tuning.

1 Introduction

Bridges are a critical civil infrastructure enabling mobility, fostering social connectivity, and facilitating economic growth. However, their structural integrity is compromised by several factors, including harsh environmental conditions, freeze–thaw cycles, exposure to de-icing chemicals, deferred maintenance, increasing traffic volumes, and aging [1]. Preventive measures are essential to ensure these structures do not deteriorate to the point of compromising safety or requiring closures. Bridge deterioration is manifest in defects such as rebar corrosion, cracking, delamination, scaling, and spalling of concrete, with data on the type, location, and severity of these defects collected during bridge inspections. However, there are significant challenges concerning the effective utilization of bridge inspection data, as it is often scattered across unstructured textual formats in inspection reports [2, 3]. The extraction, structure, and characterization of this data are further complicated by the need for automated, data-driven analyses to support condition assessments, predictive modeling, and decision-making. These barriers hinder the development of reliable asset management tools, delaying preventive and corrective actions and jeopardizing bridge safety [4].

Natural Language Processing (NLP) offers a promising solution by enabling the extraction of critical information from unstructured textual data [5, 6]. Advances in Large Language Models (LLMs) and Generative Artificial Intelligence (GenAI) have further enhanced NLP's capabilities, making it increasingly effective in processing and analyzing complex technical information [7, 8]. These cutting-edge AI techniques can automate the extraction and "standardization" of bridge inspection data. In the present study, "standardization" refers to the systematic transformation of unstructured and heterogeneous bridge inspection data into a consistent, ready-to-use, machine-readable format, enabling better-

informed condition assessments, predictive modeling, and maintenance planning, ultimately improving the reliability and efficiency of bridge asset management. While numerous studies have proposed methods to extract and structure data from bridge inspection reports, e.g., [2, 9-13], a significant gap remains with respect to transforming this structured data into formats ready for direct use in condition assessment and predictive modeling.

Liu and El-Gohary [14] proposed a hybrid data fusion method combining named entity normalization and entropy-based numerical fusion to unify bridge deficiency data. Similarly, Liu and El-Gohary [15] introduced a spectral clustering-based approach to link records describing the same bridge components or deficiencies. Li et al. [10] developed a context-aware deep-learning framework using bi-directional long short-term memory and conditional random fields to extract defect-related information, although it is noteworthy that their approach requires further standardization to align the extracted data with predictive modeling requirements. Zhang et al. [13] combined rule-based methods with generation-based LLMs to structure bridge inspection data, but noted that, without standardization, the extracted information remains unsuitable for advanced analytics. The present study takes the next step by converting extracted bridge inspection data into fully standardized formats that directly support more accurate condition assessment and predictive modeling.

LLMs, which are trained on vast datasets, excel in tasks such as translation, summarization, and question-answering by capturing nuanced language patterns and long-range dependencies [16-18]. GenAI extends LLM capabilities to generate novel content across different modalities, including text, images, and videos, using generative adversarial networks and advanced transformer architectures [8, 19]. OpenAI's GPT models, including GPT-3.5 and GPT-4, among others, exemplify these capabilities. GPT-3.5, with 175 billion parameters [17], is capable of efficiently performing a wide range of tasks [20], while GPT-4, with significantly more parameters (estimated to be in the trillions), offers enhanced accuracy and versatility [8, 21]. These models have been applied across diverse domains, such as integrating ChatGPT with augmented reality for operation and maintenance guidance [7], and enabling BIM-based virtual assistants [22].

Nevertheless, the following gaps are identified: (1) *lack of standardization in inspection data*—existing studies focus on data extraction but fail to standardize it for applications such as condition assessment and predictive modeling; (2) *underutilized potential of GenAI models, such as GPTs, beyond their role as chatbots*—GenAI's capabilities in asset management remain untapped, with conventional NLP approaches struggling to handle the complexity and variability of inspection data; and (3) *need for fine-tuned domain-specific models*. In this regard, general-purpose GenAI models lack precision to interpret technical language in inspection reports, whereas fine-tuned, systematic models could significantly improve this process.

2 Objectives

This study introduces a novel method leveraging GPT to standardize inspection data for improved information extraction, enabling automated condition assessments and predictive modeling. This is achieved through the following sub-objectives: (1) developing unified guidelines for defect identification and classification to ensure consistency; (2) harnessing a baseline GPT using optimized prompt engineering for data standardization; (3) fine-tuning the baseline with a large volume of domain-specific bridge inspection data; and (4) conducting a comprehensive comparative analysis between the two models.

3 Developed Method

An overview of the method developed in the present study is depicted in Figure 1. Details of each step are provided in the following subsections.

3.1 Case Study and Data

This study focused on bridge decks and their structural elements for a set of concrete bridges in Québec that are among the oldest in Canada [23] and are managed by the Ministry of Transport and Sustainable Mobility of Québec (MTQ). The analysis was concentrated on principal deck elements (i.e., structural slab and exterior sides) [24]. Concrete defects, i.e., corrosion, cracking, delamination, scaling, and spalling, were evaluated. In this respect, two sets of data were collected: the first set consisting of inspection comments extracted from bridge inspection reports and the latter consisting of defect severity thresholds extracted from the MTQ *Structures Inspection Manual* [24].

The dataset, extracted from 2,255 inspection reports and comprising 9,430 data points, contained detailed attributes such as structure number, inspected elements, condition ratings, and inspector comments (which in turn contained detailed descriptions of the location, type, and intensity of concrete defects in different bridge deck elements). The data extraction process, illustrated in Figure 1, involved systematically retrieving and preparing the bridge inspection report data so that it could be transformed from unstructured textual data into a structured format for further analysis. The reports, typically stored across multiple sources, were automatically retrieved and prepared for data extraction. These reports then underwent automated data extraction using a rule-based approach integrating Docparser and Python via a REST API. This process involved dividing reports into manageable segments, annotating samples, creating extraction rules, and iteratively verifying and

refining the extracted data for accuracy. For further details on the data collection and extraction processes—along with examples of the unstructured inspection reports, the structure of condition data stored in these reports, and the structure of the data input/output as part of the standardization process—the interested reader may refer to the authors' previous study [12]. Data concerning defect severity thresholds (Table 1) was extracted from the aforementioned MTQ *Structures Inspection Manual* [24]. This manual categorizes severity levels from slight to severe based on standardized criteria, guiding the classification of defects. All visualizations and examples provided in this paper were translated from the original language of the collected data (i.e., French) to English.

To illustrate the developed method, the following inspection comment can be considered: "*Delamination and spalling, cracks less than 0.8 mm, moderate scaling, and rust traces and deposits*." According to Table 1, the following classifications can be made:

- *Corrosion*: rust traces are classified as severity level S_2.
- *Cracking*: cracking $<$ 0.8 mm in width is classified as S_2.
- *Delamination*: delamination is classified as S_3.
- *Scaling*: moderate scaling is classified as S_2.
- *Spalling*: spalling is classified as S_3.

This example highlights the process of standardizing inspection comments by mapping the identified defects to their respective severity levels, ensuring systematic characterization for future utilization in condition assessment and predictive modeling, among other uses. However, applying this process to a large dataset poses significant challenges due to the variability in language styles and terminology. For example, the same given observation could be expressed differently (i.e., in different formats or phrasing) by different inspectors. As such, the models must effectively manage linguistic variability while ensuring consistent classification across the dataset.

The dataset, comprising 9,430 records, was divided as follows. A subset of 10% (947 records) was manually annotated and standardized to create the training/testing set, which was used to fine-tune the base model and evaluate model performance. This training/testing set was then further divided, with 30% (285 records) allocated to fine-tuning (focusing on refining the prompt engineering and optimizing the model's responses), and the remaining 70% (662 records) forming the testing subset for assessing the comparative performance of the models. The remaining of the dataset (90%: 8,483 records) was used as the generalization set to validate the best-performing model on a broader scale, ensuring its robustness, generalizability, and practical applicability in real-world scenarios.

3.2 Prompt Engineering

This study leveraged the power of GPT-3.5 for processing and standardizing unstructured bridge inspection data. It should be noted that this specific model was selected due to its cost-effectiveness and its having been recommended by OpenAI as the most suitable for fine-tuning, at the time of conducting the experiments in this study [25]. Prompt engineering, it should be noted, is a foundational step in guiding LLMs to produce accurate and contextually relevant responses. The design of prompts is essential when dealing with complex datasets such as bridge inspection comments. In this study, the prompt engineering process, illustrated in Figure 2(a) and Figure 2(d), integrated the following components:

- *Defect types*: defining defect categories such as corrosion, cracking, delamination, scaling, and spalling ensures that the model effectively identifies and categorizes the specific issues to be analyzed.
- *Severity levels*: standardizing defect severity levels into categories, i.e., slight, moderate, significant, and severe, enables consistent evaluation across inspection records.
- *Classification rules and thresholds*: establishing clear rules and thresholds provides a framework for consistent defect classification, ensuring the model adheres to predefined criteria.
- *Examples*: providing practical examples demonstrates how rules and thresholds are applied, enhancing the model's understanding and performance.
- *Goal definition*: specifying the purpose of the analysis (e.g., standardizing defect severity classification) makes the model's output focused on actionable insights.
- *Input and output formats*: clearly defining input and output structures ensures that the model correctly interprets raw data and generates structured, consistent outputs for downstream analysis.

These components were carefully combined to create the "*prompt base*," a foundational structure containing essential instructions and context. Inspection batch records were dynamically appended to the prompt base, creating a complete set of prompts to guide the model's analysis. This modular design, it should be noted, ensures flexibility and facilitates updates, as text files containing prompt components can be modified directly without requiring programming expertise. In this manner, the structured prompts streamline the workflow and maintain consistency.

3.3 Fine-tuning

Fine-tuning is a crucial process that customizes the LLM to enhance its performance for domain-specific tasks. In this study, the fine-tuning process was conducted using *GPT-3.5-turbo-0125*, a model optimized for such applications. The fine-tuning workflow, presented in Figure 2(b), involved the following steps:

- *Dataset preparation*: fine-tuning datasets were prepared with a training set (90%) and a validation set

(10%) to ensure effective learning and generalization; each dataset contained examples reflecting expected inputs and responses, formatted to align with the fine-tuning requirements and designed prompts.

- *Training configuration*: the model was trained (fine-tuned) using 1,399,893 tokens over three epochs, with hyperparameters such as a batch size of 1, a learning rate multiplier of 2, and a constant seed value to ensure reproducibility.
- *Performance metrics*: training and validation losses, along with training accuracy, were tracked to evaluate the model's learning and generalization capabilities.
- *Checkpoints*: intermediate checkpoints were created during training to monitor performance and prevent overfitting. These checkpoints enabled flexibility in evaluating the model at various stages of training.
- *Evaluation and iteration*: the model's outputs were evaluated against ground truth data.

Following prompt engineering and fine-tuning, the performance of both the baseline model GPT-3.5-turbo (BL-GPT) and the fine-tuned model (FT-GPT) was validated through an iterative testing process. Figure 2(c) illustrates the workflow, which involved dividing datasets into samples and batches, incorporating delays to comply with API rate limits, and analyzing performance against ground truth data. The standardized outputs were then evaluated, and the process was refined iteratively until the desired performance criteria were met.

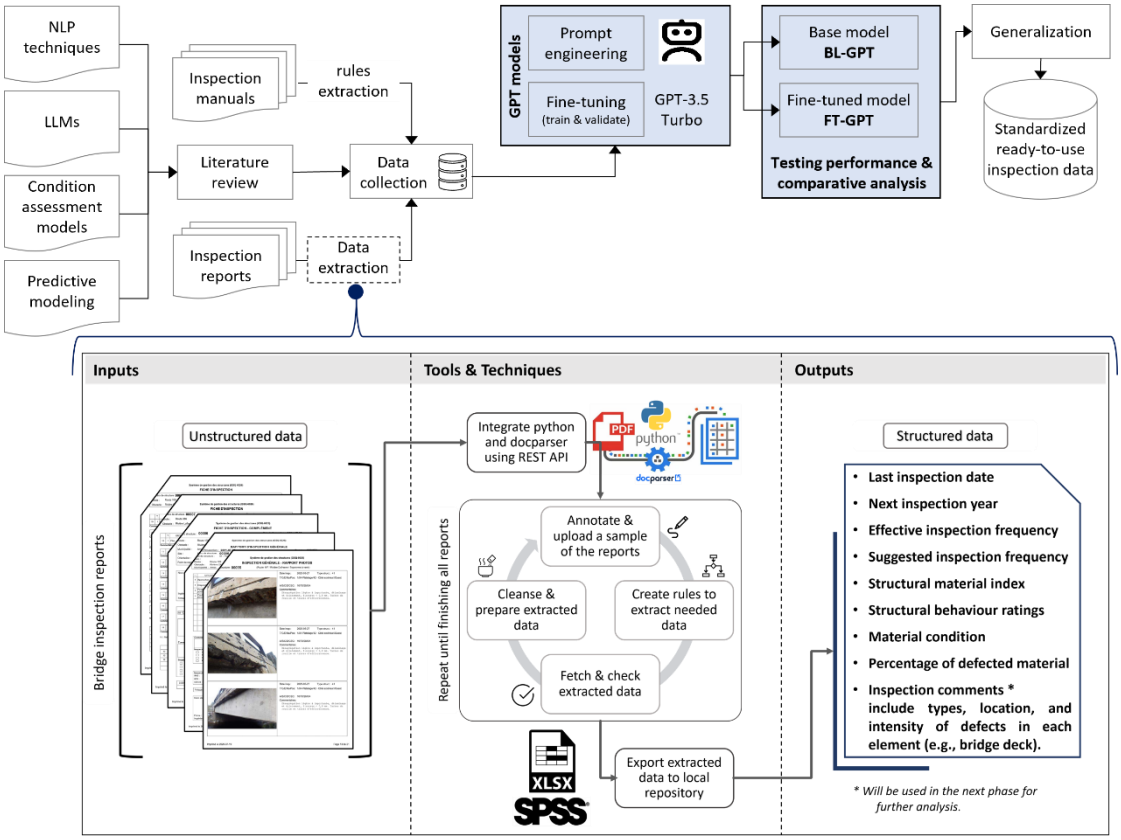

Figure 1. Overview of the developed method

Table 1. Thresholds for defect severity in bridge inspections [24]

	Severity level			
Defect	Slight (S_1)	Moderate (S_2)	Significant (S_3)	Severe (S_4)
Corrosion	Rust-free	Rusty concrete surface or reinforcement without cross-section reduction	Visible rusty reinforcement and reduced cross-section up to 30%	Very rusty reinforcement with greater than 30% reduction in cross-section
Cracking	—	< 0.8 mm	0.8–3.0 mm	> 3.0 mm
Delamination	—	—	Presence	—
Scaling	< 25 mm	25–50 mm	50–100 mm	> 100 mm
Spalling	—	—	Presence	—

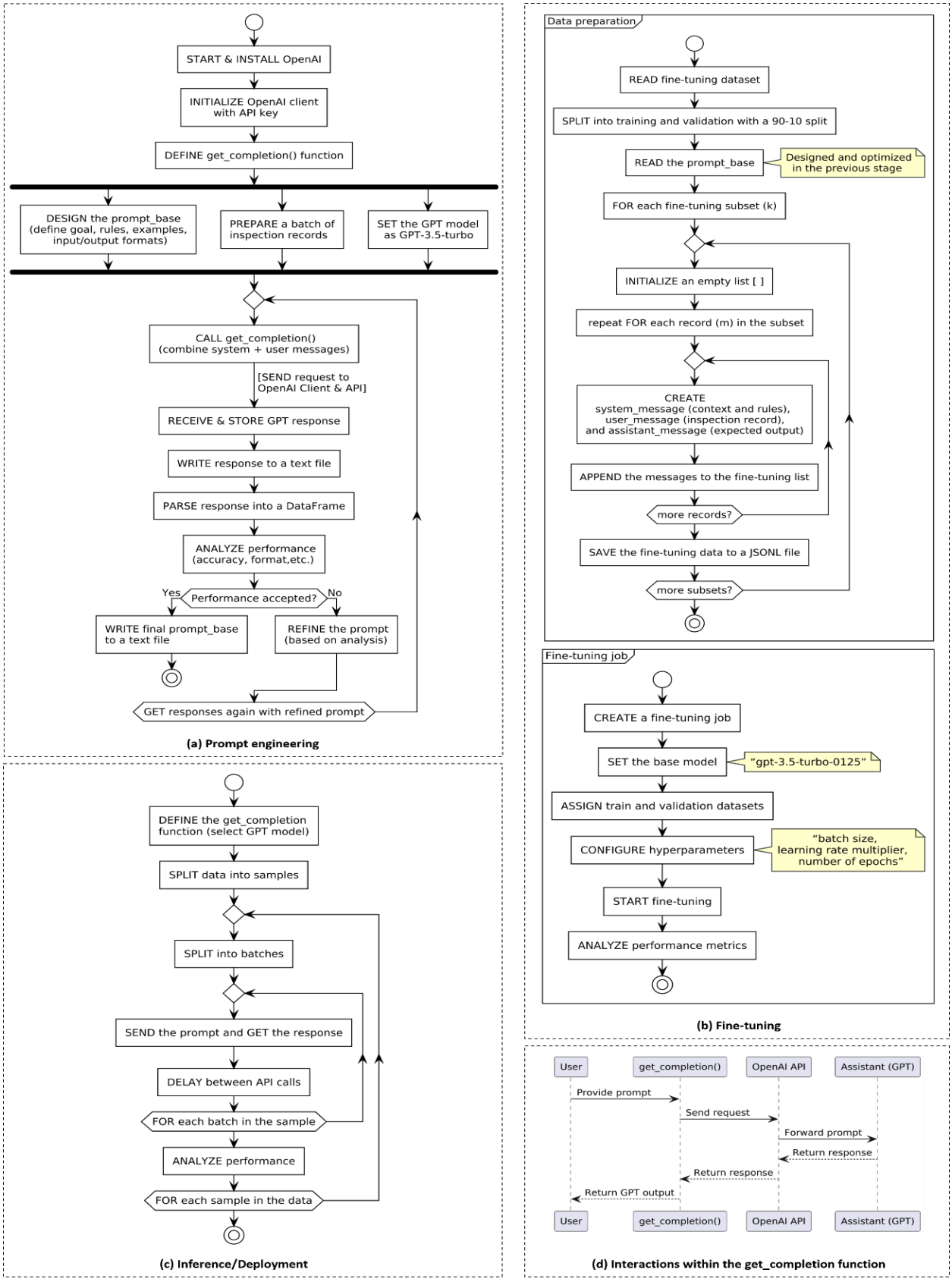

Figure 2. Workflow for GPT-based bridge data processing

4 Results and Discussion

The analysis of the 9,430 inspection records identified frequently occurring terms such as "cracking," "spalling," and "delamination." Figure 3 illustrates the structure of the input data to the BL-GPT and FT-GPT, along with the structure of the output data from the fine-tuned model. The BL-GPT and FT-GPT were tested on the 662-record test dataset, followed by a comprehensive comparative analysis of their performance in identifying and categorizing concrete bridge deck defects. The evaluation focused on accuracy, precision, recall, and F1 score across defect types, as represented by Eqs. (1) – (4), respectively, where TP, TN, FP, and FN are defined as True Positive, True Negative, False Positive, and False Negative, respectively. For the multi-class problem at hand, precision, recall, and F1 score were computed individually for each class and then aggregated using the weighted average method.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F1\ score = \frac{2TP}{2TP + FP + FN} \quad (4)$$

As shown in Table 2, FT-GPT consistently outperformed BL-GPT in the detection and classification of all defect types, achieving exceptional performance and reliability. BL-GPT exhibited particularly poor performance in detecting and characterizing spalling, achieving only 61.63% accuracy and a 54.50% F1 score, substantially lower than FT-GPT's 98.64% in both metrics. The lower performance metrics of the base model is attributable to several factors:

- *Complexity of inspection comments*: Inspection comments often include multiple defect descriptions and measurements, introducing complexity and ambiguity that result in misinterpretation of context.
- *Insufficient training data and generalized knowledge*: The base model lacks exposure to a diverse and specialized dataset, limiting its ability to generalize effectively to complex inspection comments. Although it is pre-trained on extensive general-purpose data, its generalized knowledge is insufficient for the technical and domain-specific language used in bridge inspection reports. As such, without fine-tuning on domain-specific data, the model is unable to fully adapt to the terminology, phrasing, and contextual nuances unique to bridge inspection comments, resulting in less accurate and consistent standardization.
- *Stochastic nature*: Operating probabilistically with billions of parameters, the base model may produce varying outputs even with the same input, leading to inconsistencies in defect classification and severity assessment.
- *Language mismatch and translation nuances*: The process of translating between languages—in this study, the original data was in French, while the prompts and interactions with the model were designed in English—can result in the loss of subtle meanings and technical nuances.

The results of this study demonstrate that fine-tuning is as an efficient solution to overcome the limitations outlined above. By adapting the model to domain-specific data, fine-tuning enhances its ability to accurately interpret terminology, phrasing, and contextual details unique to bridge inspections. This process mitigates issues related to contextual misinterpretation, complexity, and insufficient domain knowledge. Furthermore, fine-tuning improves output reliability and stability, ensuring accurate defect classification and severity assessment. It also strengthens the model's capacity to handle multilingual inputs, maintaining the integrity of technical terms during translation and standardization. The fine-tuned model offers several advantages over conventional NLP techniques, particularly in handling complex and unstructured textual data. It reduces the need for extensive programming and manual feature crafting, allowing engineers to interact with the model and scale it to their specific tasks using their linguistic patterns. Through the use of a well-designed prompt, the model eliminates the need for complex coding expertise, making it accessible and adaptable to various practical applications.

The best-performing model, FT-GPT, was applied to the generalization dataset, which comprised 8,483 records (i.e., 90% of the overall dataset). This step involved evaluating the model's robustness, generalization capabilities, and practical applicability. The dataset was enriched (i.e., through generalization) to include standardized inspection narratives, the severities of different types of reinforced concrete bridge deck defects presented in a consistent, standardized format, and associated condition ratings for each element in these bridge decks. This data asset can be utilized in future studies on the asset management of bridges. As a proof of concept, the data has already been used in developing data-driven condition assessment models to evaluate the condition of bridge deck elements and overall deck condition based on the identified defect severities. (For further details, the interested reader may refer to the authors' previous work [12].)

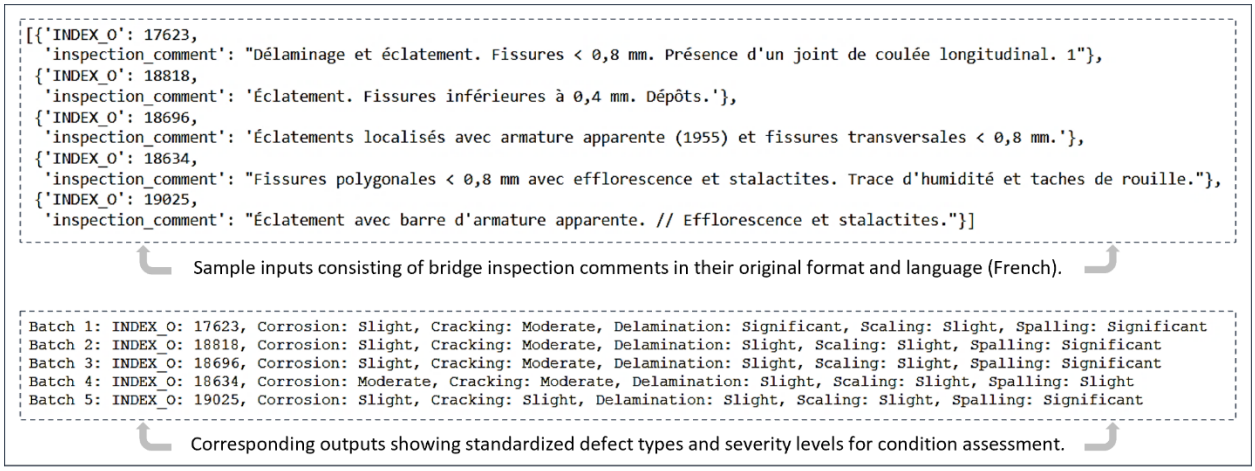

Figure 3. Structure of inputs and outputs of the fine-tuned model

Table 2. Performance comparison of BL-GPT and FT-GPT models in detecting and standardizing concrete defects

	Model	BL-GPT				FT-GPT			
	Performance	Accuracy	Precision	Recall	F1 score	Accuracy	Precision	Recall	F1 score
Defect	Corrosion	88.22%	88.43%	88.22%	87.44%	98.79%	98.99%	98.79%	98.84%
	Cracking	92.90%	94.15%	92.90%	93.35%	98.64%	98.68%	98.64%	98.58%
	Delamination	96.37%	97.41%	96.37%	96.88%	99.09%	99.10%	99.09%	99.09%
	Scaling	91.24%	88.53%	91.24%	88.37%	98.64%	98.78%	98.64%	98.64%
	Spalling	61.63%	78.09%	61.63%	54.50%	98.64%	98.64%	98.64%	98.64%

As of May, 2024, i.e., at the time of conducting the experiments for this study, the implementation cost and processing time for analyzing 100 batches of inspection comments were evaluated. The BL-GPT model incurred a total cost of ~$0.09, processing ~167,238 tokens, consisting of ~161,560 input tokens and ~5,678 output tokens, with an average processing time of approximately 7 min. In contrast, the fine-tuned GPT model had a higher total cost of ~$0.52, with ~3,449 output tokens and a slightly reduced processing time of approximately 6.5 min. This highlights that the fine-tuned model generates more concise responses while incurring a lower processing time. The fine-tuning process itself (with the number of tokens and parameters discussed in Section 3.3) incurred a total cost of ~$11.19.

5 Summary and Concluding Remarks

This study represents a significant step toward leveraging GenAI models for asset management and enhancing the utilization of bridge inspection data. The process focuses on detecting and categorizing defects in reinforced concrete bridge decks—including corrosion, cracking, delamination, scaling, and spalling—into defined severity levels. LLM-based GPT-3.5 was harnessed as a baseline model. This model was fine-tuned with domain-specific data to enhance its ability to process complex technical information from bridge inspection reports. A robust prompt engineering process was developed to effectively capture and standardize defect severity levels. The fine-tuned model demonstrated superior performance and was found to be reliable in standardizing inspection comments. The contributions of this study include (1) introducing a GPT-based method for improved information extraction from and standardizing bridge inspection data to support data-driven decision-making; (2) demonstrating the effective use of GenAI in infrastructure management; (3) fine-tuning LLMs to improve performance, stability, and reliability; (4) employing robust prompt engineering in asset management; and (5) providing a standardized dataset for utilization in future work. While this study focused on defects in concrete bridge decks, the methods are extendable to other infrastructure systems, paving the way for improved infrastructure management more broadly. It should be noted that further investigations are needed to evaluate the implementation costs associated with these models. Future research should also explore the performance of newer models, such as GPT-4o and o1, as well as open-source alternatives, such as Llama and Gemini.

References

[1] A. Omar and O. Moselhi. Condition monitoring of reinforced concrete bridge decks: Current practices and future perspectives. *Current Trends in Civil & Structural Engineering*, 8(4), 2022.

[2] K. Liu and N. El-Gohary. Ontology-based semi-supervised conditional random fields for automated information extraction from bridge inspection reports. *Automation in Construction*, 81:313–327, 2017.

[3] R. Li, T. Mo, J. Yang, D. Li, S. Jiang, and D. Wang. Bridge inspection named entity recognition via BERT and lexicon augmented machine reading comprehension neural model. *Advanced Engineering Informatics*, 50:101416, 2021.

[4] A. Omar and O. Moselhi. An integrated approach for automated acquisition of bridge data and deficiency evaluation, In *Proceedings of the 40th International Symposium on Automation and Robotics in Construction (ISARC)*, Vol. 40, IAARC, Chennai, India, 2023.

[5] X. Chen, H. Xie, and X. Tao. Vision, status, and research topics of Natural Language Processing. *Natural Language Processing Journal*, 1:100001, 2022.

[6] J. Just. Natural language processing for innovation search – Reviewing an emerging non-human innovation intermediary. *Technovation*, 129:102883, 2024.

[7] F. Xu, T. Nguyen, and J. Du. Augmented reality for maintenance tasks with ChatGPT for automated text-to-action. *Journal of Construction Engineering and Management*, 150(4):04024015, 2024.

[8] P. Ghimire, K. Kim, and M. Acharya. Opportunities and challenges of Generative AI in construction industry: Focusing on adoption of text-based models. *Buildings*, 14(1):220, 2024.

[9] K. Liu and N. El-Gohary. Semantic neural network ensemble for automated dependency relation extraction from bridge inspection reports. *Journal of Computing in Civil Engineering*, 35(4):04021007, 2021.

[10] T. Li, M. Alipour, and D. K. Harris. Context-aware sequence labeling for condition information extraction from historical bridge inspection reports. *Advanced Engineering Informatics*, 49:101333, 2021.

[11] S. Moon, S. Chung, and S. Chi. Bridge damage recognition from inspection reports using NER based on recurrent neural network with active learning. *Journal of Performance of Constructed Facilities*, 34(6):04020119, 2020.

[12] A. Omar and O. Moselhi. Automated data-driven condition assessment method for concrete bridges. *Automation in Construction*, 167:105706, 2024.

[13] C. Zhang, X. Lei, Y. Xia, and L. Sun. Automatic bridge inspection database construction through hybrid information extraction and large language models. *Developments in the Built Environment*, 20:100549, 2024.

[14] K. Liu and N. El-Gohary. Fusing data extracted from bridge inspection reports for enhanced data-driven bridge deterioration prediction: A hybrid data fusion method. *Journal of Computing in Civil Engineering,* 34(6):04020047, 2020.

[15] K. Liu and N. El-Gohary. Improved similarity assessment and spectral clustering for unsupervised linking of data extracted from bridge inspection reports. *Advanced Engineering Informatics*, 51:101496, 2022.

[16] S. Banerjee, C. M. Potts, A. H. Jhala, and E. J. Jaselskis. Developing a construction domain–specific artificial intelligence language model for NCDOT's CLEAR Program to promote organizational innovation and institutional knowledge. *Journal of Computing in Civil Engineering*, 37(3):04023007, 2023.

[17] T. Nguyen-Mau, A.-C. Le, D.-H. Pham, and V.-N. Huynh. An information fusion based approach to context-based fine-tuning of GPT models. *Information Fusion*, 104:102202, 2024.

[18] S. Moon, S. Chi, and S.-B. Im. Automated detection of contractual risk clauses from construction specifications using bidirectional encoder representations from transformers (BERT). *Automation in Construction*, 142:104465, 2022.

[19] S. Chung, S. Moon, J. Kim, J. Kim, S. Lim, and S. Chi. Comparing natural language processing (NLP) applications in construction and computer science using preferred reporting items for systematic reviews (PRISMA). *Automation in Construction*, 154:105020, 2023.

[20] OpenAi. OpenAI platform documentation: Models. On-line: https://platform.openai.com/docs/models, Accessed: 11 June 2024.

[21] OpenAi. GPT-4. On-line: https://openai.com/gpt-4/, Accessed: 11 June 2024.

[22] J. Zheng and M. Fischer. Dynamic prompt-based virtual assistant framework for BIM information search. *Automation in Construction*, 155:105067, 2023.

[23] Statistics Canada. Age of public infrastructure: a provincial perspective. On-line: https://www150.statcan.gc.ca/n1/pub/11-621-m/11-621-m2008067-eng.htm, Accessed: 8 March 2022.

[24] Ministry of Transport and Sustainable Mobility of Québec. *Structures Inspection Manual*. Government of Québec, National Library and Archives of Québec, 2017.

[25] OpenAi. Fine-tuning guide. On-line: https://platform.openai.com/docs/guides/fine-tuning, Accessed: 11 June 2024.