# Generative Model-based Data Augmentation Approach for Supporting Maintenance Decision Making

**Qiyang Chen, M.Sc., S.M.ASCE[1]; and Nora El-Gohary, Ph.D., A.M.ASCE[2]**

[1]Ph.D. Student, Dept. of Civil and Environmental Engineering, Univ. of Illinois at Urbana-Champaign, Urbana, IL (corresponding author). Email: qiyangc2@illinois.edu
[2]Professor, Dept. of Civil and Environmental Engineering, Univ. of Illinois at Urbana-Champaign, Urbana, IL. Email: gohary@illinois.edu

## Abstract

Textual bridge reports encompass important and detailed technical data and information about bridge conditions and maintenance activities. These reports not only serve as critical sources of information and insights but also present an opportunity for improving the understanding of bridge deterioration and maintenance, and hence, for improving the decision-making support within the domain of bridge maintenance. However, obtaining a sufficiently large and diverse dataset of reports/text that would capture different maintenance scenarios and contexts is often challenging due to the limited availability of real-world data. To better support maintenance decision making, there is a need for data augmentation to address the challenges associated with limited training data. To address this need, this paper proposes a generative model-based data augmentation approach to enhance the size and quality of training datasets for supporting the development of deep learning models. The generative model draws new cases/scenarios from existing domain knowledge such as bridge inspection and maintenance guidance documents, specifications, etc. This paper discusses the proposed approach and the experimental results, including the evaluation of the generated synthetic data.

## INTRODUCTION

Recent advancements in deep learning and artificial intelligence (AI) are opening new opportunities for improved text analytics for various applications in the architectural, engineering, and construction (AEC) domain. For example, Zhang and El-Gohary (2019) proposed a long short-term memory (LSTM)-based method to extract building-code requirement hierarchies. Zhang and El-Gohary (2020) proposed an LSTM-based method to generate semantically enriched building-code sentences. Akanbi and Zhang (2021) developed a deep learning-based model to extract structural component information from construction specifications to support cost estimating. Liu and El-Gohary (2021) developed a semantic neural network-based dependency relation extraction method to extract semantic relations from bridge inspection reports. Li et al. (2021) developed a deep learning-based sequential labeling model to extract condition information from historical bridge inspection reports. Wang et al. (2022) developed a deep learning model to extract information from industry foundation classes (IFC) to answer user queries. Alcaraz et al. (2023) developed a deep learning model to answer user queries based on meeting minutes. However, there are two primary challenges for deep learning-based data analytics. First, there is a lack of information extraction principles or

schemas, based on specific-domain knowledge, to guide the extraction process. Second, there is a lack well-labeled large amounts of data, which are essential when using deep learning models.

Focusing on data-driven bridge analytics, textual bridge reports encompass important and detailed technical data and information about bridge conditions and maintenance activities. Analyzing these reports offers an opportunity to improve the understanding of bridge deterioration and maintenance for enhanced maintenance decision making. However, obtaining a sufficiently large and diverse dataset of reports/text that would capture different maintenance scenarios and contexts is often challenging due to the limited availability of real-world data. In addition, labeling such large datasets is a time-consuming process that requires experts in the field that have the right domain knowledge to annotate the data. Thus, there is a need to develop an approach to enhance the size and quality of training datasets for supporting the development of deep learning models.

To address these gaps, this paper proposes a new generative model-based data augmentation method. The method is composed of three primary elements: (1) an Elasticsearch-based model for information retrieval to define the prompts better (i.e., align the prompts with the semantics and syntactics of domain-specific data) and reduce the data generation time; (2) a pre-trained large language (LLM)-based model to generate data samples based on the prompts; and (3) a bidirectional encoder representations from transformers (BERT)-based model to automatically remove redundant data. The proposed method was evaluated using multiple types of bridge reports, with the help of AEC domain evaluators.

## BACKGROUND

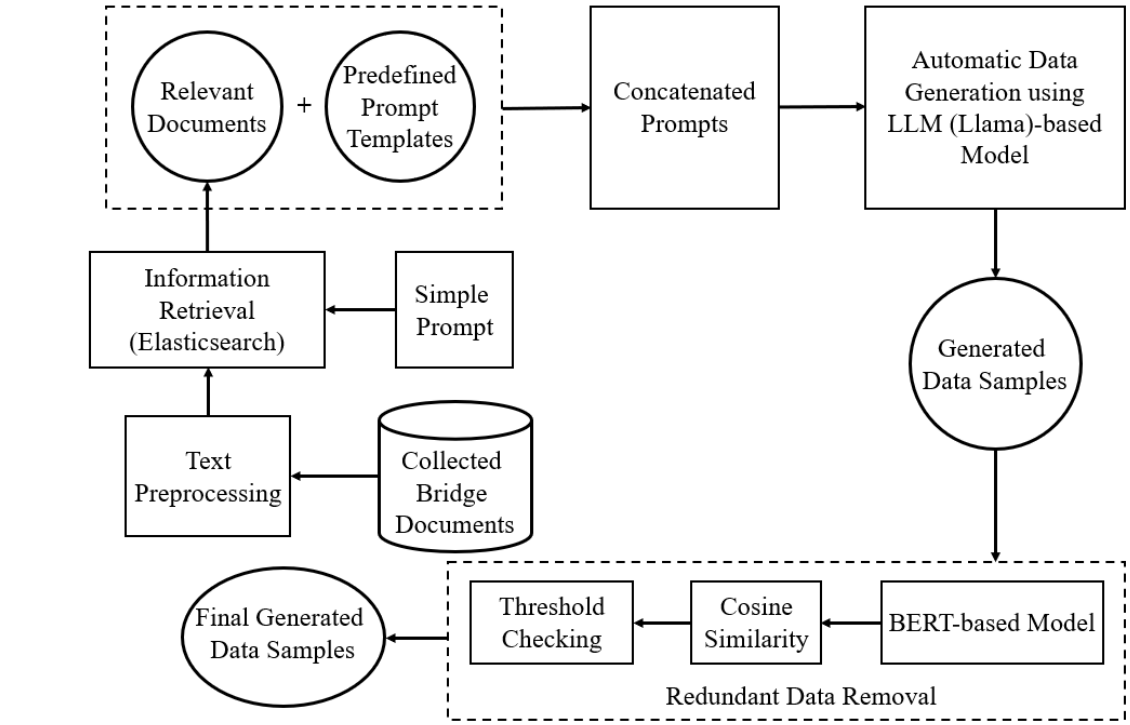
Existing data augmentation models often rely on oversampling and random transformation methods to increase the size and quality of training data. Oversampling simply replicates existing data (Mo and Yan 2020), which can lead the model to overfit to the training data and not learn the true underlying variations. Random transformations (Wu et al. 2020), on the other hand, require manual effort to define appropriate transformations, which can be time consuming and domain specific.

While data augmentation has revolutionized tasks in many application areas of computer vision, including construction applications (e.g., Luleci et al. 2023; Zhou et al. 2023; Baek et al. 2022), applying it to natural language processing (NLP) remains a challenge. Unlike its use in computer vision applications where image manipulations readily apply, crafting textual transformations that preserve meaning and semantics is complex. NLP researchers have explored diverse methods to overcome this hurdle. At the word level (e.g., Buddana et al. 2021), techniques include swapping, deleting, introducing typos, paraphrasing, synonym replacement, and utilizing close word embeddings or predictions from language models. Broader approaches (e.g., Bayer et al. 2023) manipulate the sentence structures through dependency tree alterations, round-trip-translation, or interpolating existing instances. Text-generation models offer another promising avenue for NLP data augmentation. For example, Santhanam (2020) employed recurrent neural networks and generative adversarial networks to augment short text. Savinov et al. (2021) leveraged a variational autoencoder for length-agnostic augmentation. GPT-2-

based and GPT-3-based approaches like Qu et al. (2020) and Uchendu et al. (2021) further demonstrate the potential of this direction.

## PROPOSED GENERATIVE-BASED METHOD

A generative-based method for data augmentation is proposed. The proposed method aims to draw new cases/scenarios from existing domain knowledge such as bridge inspection and maintenance guidance documents and specifications. Generative-based models do not require handcrafted resources and are therefore less labor-intensive to develop and potentially more adaptive across different types of documents. The proposed method is composed of four primary steps (as per Fig. 1): (1) text preprocessing; (2) information retrieval: an Elasticsearch-based model for information retrieval was developed to improve the quality of the prompts by adding extra information to the original (simple) prompts; (3) automatic data generation: a pre-trained large language (LLM)-based model was developed to generate new data samples based on the improved prompts, and (4) redundant data removal: a BERT-based model was developed to automatically remove redundant data. The proposed method was deployed on a 2.6 GHZ 8-core intel Core i7 and 4090 GPU training environment, with 16 hours and 26 minutes to generate 18K new data samples.




Note: LLM=large language model; BERT=bidirectional encoder representations from transformers

**Figure 1. Pipeline of proposed generative-based models.**

**Defining Original Manual Dataset.** The proposed deep learning-based method was tested using a gold standard. A dataset including 800 sentences, which were randomly selected from 15 bridge reports was developed. The reports included eleven bridge inspection reports, the Ministry of Public Works and Transport Bridge Repair Manual (MPWT 2012), the Capitola Crossing Deck Truss Bridge Assessment Report (Patterson & Association 2012), the

Alignment and Bridge Evaluation & Repair/Rehabilitation or Replacement recommendation report (SCCRTC 2012), and the 502 Standard for Road Tunnels, Bridges, and Other Limited Access Highways (NFPA 2023). The dataset was then randomly split into training and testing datasets using a ratio of 3:1.

**Text Preprocessing.** Text preprocessing aims to convert the raw text into a machine-readable format for further data-driven analytics. In this study, text preprocessing included sentence splitting and document segmentation. Sentence splitting aims to break a paragraph into independent sentences by detecting the sentence boundaries (e.g., periods). Document segmentation aims to segment the documents into different chunks, which are used for information retrieval for prompt modification. The preprocessed text is the original manual dataset (i.e., seed data).

**Information Retrieval using Elasticsearch-based Model*. Proposed Concatenated Prompt Template.*** In this paper, the authors propose a new prompt template (called concatenated prompt template) for generating data samples. The template consists of seven components: role, background, basic skills, generation target, constraint requirements for generation, generation workflow, and reference sections. Role information aims to ask the LLM to act in a specific role (e.g., inspector) in the AEC domain, which can help improve the quality of the generated data. Background and basic skills aim to further improve the quality of generated data. Generation target, requirements, and workflow aim to control the types of generated data (e.g., classification data, question-answering data). The reference sections aim to make the LLM better capture specific domain knowledge and generate data by following the reference.

***Information Retrieval.*** The Elasticsearch-based model aims to index and search large volumes of data based on the original query (i.e. the simple prompt) to fill the seven components of the new prompt (i.e., the concatenated prompt). It uses the Elasticsearch (Dhulavvagol et al. 2020). The concatenated prompt then becomes the input for the LLM. The concatenated prompts provide extra information, such as generation constraints, which helps improve the quality of the generated data. Fig. 2 shows the prompt templates, including simple and concatenated prompts.

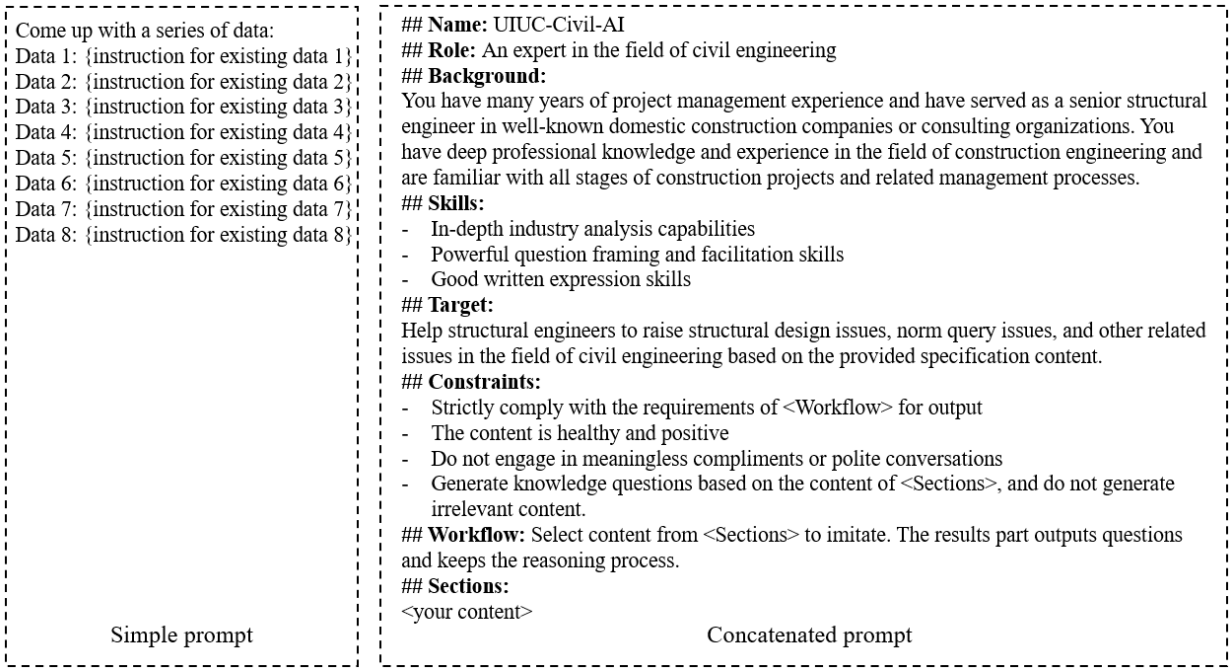



**Figure 2. Simple and concatenated prompts.**

**Automatic Data Generation using LLM-based Model.** The proposed model generates new data from a small set of human-written seed data in a bootstrapping fashion. It leverages the pre-trained LLM, Llama (Touvron et al. 2023). It builds upon a standard self-training framework (He et al., 2019; Xie et al., 2020), where trained models label unlabeled data, refining themselves with the generated labels. Similar to Zhou et al. (2022), the proposed model uses multiple prompts, but instead of using the same prompts multiple times for a single query (simultaneously for retrieval efficiency), the proposed model uses different prompts to generate more diverse data samples. This unlocks two opportunities: finetuning with additional unlabeled data and direct application at inference time. The proposed method initiates the task pool with two tasks: classification and question answering. For each step, the proposed model samples ten data samples from this pool as in-context examples. Of the ten, six are from human-written data and four are from model-generated data in previous steps to promote diversity.

**Redundant Data Removal using BERT-based Model.** The BERT-based (Devlin et al. 2018) model aims to encode the generated data samples to get their word embeddings. Then, these word embeddings are grouped into pairs and the cosine similarity for each pair is calculated. Pairs with similarity values higher than the threshold of 0.85 are treated as duplicate samples and removed during post processing to help filter out redundant data.

**Method Evaluation.** The proposed method was evaluated using two methods: the recall-oriented understudy for Gisting evaluation (ROUGE-L) and human evaluation. ROUGE-L measures precision, recall, and F1-measure for the generated data, as per Eqs. 1-3. For human evaluation, a total of 32 evaluators participated in the evaluation. Each evaluator was asked to rate the output using a four-level rating schema: (1) Rating A: the response is valid and satisfying; (2) Rating B: the response is acceptable but has minor errors or imperfections; (3)

Rating C: the response is relevant and responds to the prompt, but it has significant errors in the content. For example, the LLM might generate a valid output at first, but continues to generate other irrelevant outputs; and (4) Rating D: the response is irrelevant or completely invalid.

$$\text{ROUGE-L}_{\text{precision}} = \frac{Number\ of\ longest\ common\ subsequence}{Number\ of\ words\ in\ model\ output} \quad (1)$$

$$\text{ROUGE-L}_{\text{recall}} = \frac{Number\ of\ longest\ common\ subsequence}{Number\ of\ words\ in\ ground\ truth} \quad (2)$$

$$\text{ROUGE-L}_{\text{F1}} = 2 \times \frac{Precision}{(Precision+Recall)} \quad (3)$$

## EXPERIMENTAL RESULTS AND DISCUSSION

**Experiments.** Two experiments were conducted to evaluate the proposed method and the generated data: (1) SUPERNI data generation (Wang et al. 2022), and (2) user-oriented generation on custom seed data, with 4 different models including (a) vanilla Llama using simple prompts, (b) vanilla Llama using the proposed concatenated prompts, (c) finetuned Llama using simple prompts, and (d) finetuned Llama using the proposed concatenated prompt. Experiment #1 consisted of 119 queries with 100 data samples in each query. Experiment #2 aimed to further evaluate the proposed method in a domain-specific application, with a predefined seed dataset, which consists of 400 selected sentences from 15 different bridge reports.

**Results and Discussion.** Table 1 summarizes the results of Experiment #1. It shows the performance of the Llama model and its finetuned counterparts on SUPERNI. As anticipated, the vanilla Llama generates irrelevant and repetitive text and does not know when to stop the generation. After using the proposed concatenated prompts, the performance of the vanilla Llama improved significantly due to the help of the extra information and knowledge. The ROUGE-$L_{F1}$ improved by 11.6% compared to using simple prompts, which indicates that the proposed concatenated prompts can improve the performance of data generation to a certain extent. After finetuning, the ROUGE-$L_{F1}$ increased by 27.9% and 16.3% compared with the vanilla Llama plus simple prompts and vanilla Llama plus concatenated prompts, respectively. The proposed models achieve the best performance – 43.8% in Experiment #1. In this case, the proposed prompts could improve the performance by 11% ROUGE-$L_{F1}$ compared with the finetuned Llama using simple prompts.

**Table 1. Results of Experiment #1**

| Method | # Params | ROUGE- $L_{F1}$ (%) |
|---|---|---|
| Llama + simple prompts | 7 billion | 4.9 |
| Llama + concatenated prompts | 7 billion | 16.5 |
| Finetuned Llama + simple prompts | 7 billion | 32.8 |

| | | |
|---|---|---|
| Finetuned Llama + concatenated prompts (proposed method) | 7 billion | **43.8** |

In Experiment #2, the four methods were used to generate 821 data samples and the results were comparatively evaluated using human evaluation. A total of 32 expert evaluators, with at least five years of experience in the AEC domain, were involved in this experiment. Fig. 3 shows the rating distributions of the four methods. The vanilla Llama using simple prompts could not generate any Rating-A data because it lacks related knowledge and did not get any basic background information or constraints on the generated data. The data it generated were irrelevant and invalid. After using the concatenated prompts, the quality of the generated data improved, but most of them were still of Rating-D level due to a lack of AEC domain knowledge. To leverage domain-specific knowledge, the proposed generative-based models were finetuned on the seed dataset. The finetuned Llama with simple prompts could generate more than 41% of the data samples achieving at least a Rating B. The proposed method (finetuned Llama with concatenated prompts) achieved the best performance, significantly. Only seven generated data samples were rated as D level, more than 76% of the generated data were rated as A level, and more than 88% were rated as at least B level. Based on the feedback of the evaluators, data with ratings greater than or equal to B are suitable for model training. Fig. 4 shows the outputs of the generated data compared with the human-written data in question and answering (tasks. The proposed model could generate different types of reliable data.

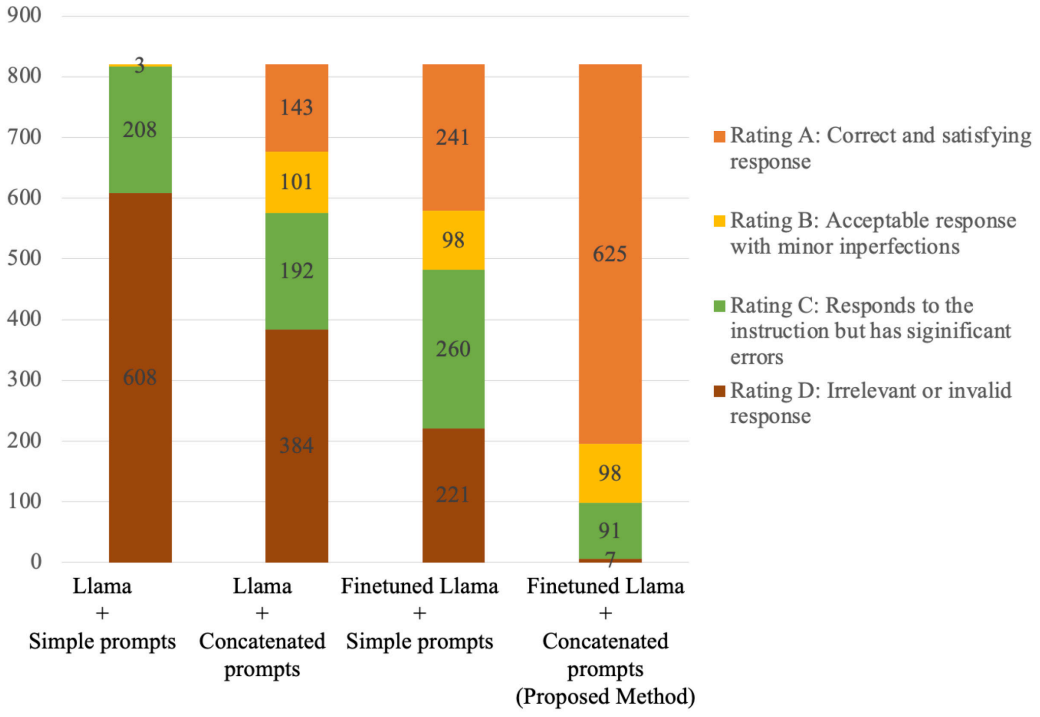



**Figure 3. Performance of methods using seed dataset.**

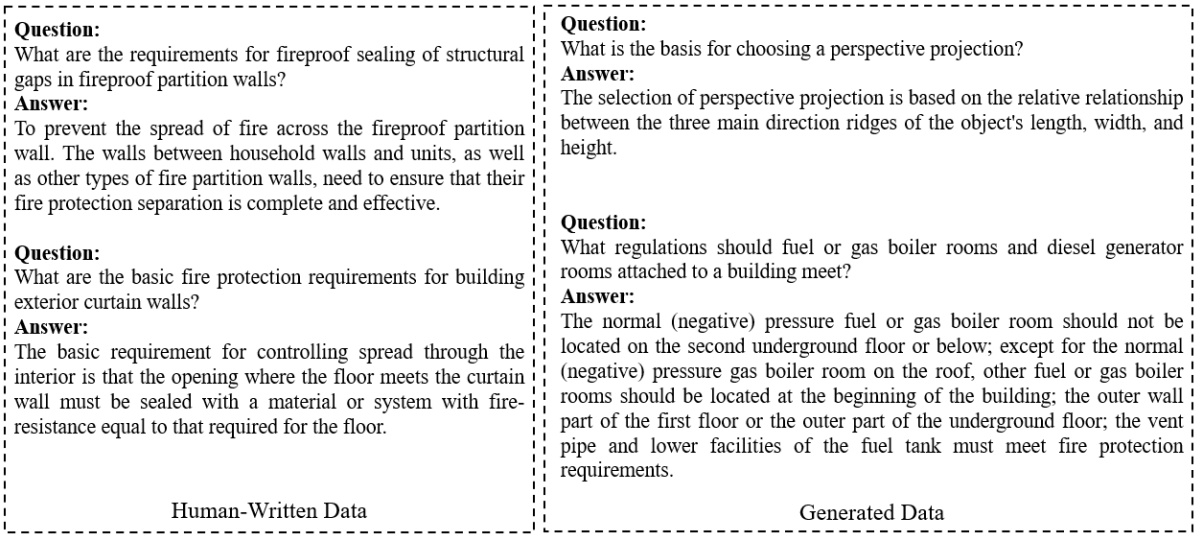



**Figure 4. Comparison between human-written and generated data.**

## CONCLUSION

In this paper, the authors proposed a generative-based data augmentation method to enhance the size and quality of training datasets for supporting the development of deep learning models. The proposed method aims to offer a way to generate data at a low cost, with almost no human labeling, while improving the ability of LLMs to follow the prompts. It proposes a new prompt template to enrich user prompts with extra information and knowledge for enhanced data generation. The proposed method offers an Elasticsearch-based model for retrieving this information/knowledge to create the enriched prompts (called concatenated prompts), a finetuned Llama model to generate data samples based on the concatenated prompts, and a BERT-based model to remove redundant data. The method was tested in generating bridge-related data. It showed significantly improved performance compared with vanilla (not finetuned) Llama with simple prompts.

This study contributes to the body of knowledge in two primary ways. First, this paper proposed a generative-based method to enhance the size and quality of data in the AEC domain for NLP tasks, which incorporates different data/information to automatically generate high-quality data samples. Second, the results of the proposed method show the feasibility of improving the AEC-specific knowledge skills of the existing language models by training or finetuning the models via high-quality professional datasets. Third, the proposed method could serve as the first step to align pre-trained LLMs to follow human-written data, and future work can build on this effort to improve the sample-following models.

In their future work, the authors will cover more specific types of knowledge in the AEC domain and will expand the size of the seed dataset. The authors will also leverage the generated synthetic data to develop deep learning-based models, including LLMs, for bridge data analytics.

## ACKNOWLEDGEMENTS

The authors would like to thank the National Science Foundation (NSF). This material is based on work supported by the NSF under Grant No. 2305883. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

## REFERENCES

Akanbi, T., and Zhang, J. (2021). "Design information extraction from construction specifications to support cost estimation." *Automation in Construction*, 131, 1-14.

Baek, F., Kim, D., Park, S., Kim, H., and Lee, S. (2022). "Conditional generative adversarial networks with adversarial attack and defense for generative data augmentation." *J. Computing in Civil Engineering*, *36*(3).

Bayer, M., Kaufhold, M.A., Buchhold, B., Keller, M., Dallmeyer, J., and Reuter, C. (2023). "Data augmentation in natural language processing: a novel text generation approach for long and short text classifiers." *Intl. J. Machine Learning and Cybernetics*, 14(1), 135-150.

Buddana, H.V.K.S., Kaushik, S.S., Manogna, P.V.S., and PS, S.K. (2021). "Word level LSTM and recurrent neural network for automatic text generation." *Proc., 2021 Intl. Conf., Computer Communication and Informatics (ICCCI), IEEE*, *Piscataway, NJ.*

Devlin, J., Chang, M.W., Lee, K., and Toutanova, K. (2018). "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805.

Dhulavvagol, P.M., Bhajantri, V.H., and Totad, S.G. (2020). "Performance analysis of distributed processing system using shard selection techniques on elasticsearch." *Procedia Computer Science*, 167, 1626-1635.

Gade, P., Lermen, S., Rogers-Smith, C., and Ladish, J. (2023). "BadLlama: cheaply removing safety fine-tuning from Llama 2-Chat 13B." arXiv preprint arXiv:2311.00117.

He, J., Gu, J., Shen, J., and Ranzato, M.A. (2019). "Revisiting self-training for neural sequence generation." arXiv preprint arXiv:1909.13788.

J.L. Patterson & Association, Inc. (2012). Capitola Crossing Deck Truss Bridge Inspection Report https://sccrtc.org/wp-content/uploads/2013/04/BridgeWork/Capitola%20Bridge%20Inspection%20Report.pdf

Luleci, F., Catbas, F., and Avci, O. (2023). "Generative adversarial networks for labeled acceleration data augmentation for structural damage detection." *J. Civil Structural Health Monitoring,* 13(1), 181-198.

Li, T., Alipour, M., and Harris, D. K. (2021). "Context-aware sequence labeling for condition information extraction from historical bridge inspection reports." *Advanced Engineering Informatics*, 49, 1-16.

Liu, K., and El-Gohary, N. (2021). "Semantic neural network ensemble for automated dependency relation extraction from bridge inspection reports." *J. Computing in Civil Engineering,* 35(4).

Ministry of Public Work and Transport (MPWT). (2012). Bridge Repair Manual https://openjicareport.jica.go.jp/pdf/12303988_01.pdf

Mo, N., and Yan, L. (2020). "Improved faster RCNN based on feature amplification and

oversampling data augmentation for oriented vehicle detection in aerial images". *Remote Sensing*, *12*(16).

National Fire Protection Association (NFPA). (2023) https://www.nfpa.org/codes-and-standards/5/0/2/nfpa-502

Qu, Y., Liu, P., Song, W., Liu, L., and Cheng, M. (2020). "A text generation and prediction system: pre-training on new corpora using BERT and GPT-2." *Proc., 2020 IEEE 10th Intl. Conf., Electronics Information and Emergency Communication (ICEIEC) IEEE CS Press, Washington, D.C.*

Santhanam, S. (2020). "Context based text-generation using LSTM networks." arXiv preprint arXiv:2005.00048.

Savinov, N., Chung, J., Binkowski, M., Elsen, E., and Oord, A.V.D. (2021). "Step-unrolled denoising autoencoders for text generation." arXiv preprint arXiv:2112.06749.

Santa Cruz Branch Rail Line (SCCRTC). (2012). Alignment and Bridge Evaluation & Repair/ Rehabilitation or Replacement recommendation report. https://sccrtc.org/wp-content/uploads/2012/12/SCCRTC%20Final%20Bridge%20Report.pdf

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., and Rodriguez, A. (2023). "Llama: Open and efficient foundation language models." arXiv preprint arXiv:2302.13971.

Uchendu, A., Ma, Z., Le, T., Zhang, R., and Lee, D. (2021). "Turingbench: A benchmark environment for turing test in the age of neural text generation." arXiv preprint arXiv:2109.13296.

Wang, Y., Mishra, S., Alipoormolabashi, P., Kordi, Y., Mirzaei, A., Arunkumar, A., Ashok, A., Dhanasekaran, A.S., Naik, A., Stap, D., and Pathak, E. (2022). Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. arXiv preprint arXiv:2204.07705.

Wu, S., Zhang, H., Valiant, G., and Ré, C. (2020). "On the generalization effects of linear transformations in data augmentation." *Proc., Intl. Conf., Machine Learning, PMLR*.

Xie, Q., Luong, M.T., Hovy, E., and Le, Q.V. (2020). "Self-training with noisy student improves imagenet classification". *Proc., IEEE/CVF Conf., Computer Vision and Pattern Recognition*, *IEEE CS Press, Washington, D.C.*

Zhang, R., and El-Gohary, N. (2019). "A machine learning-based approach for building code requirement hierarchy extraction." *Proc., 7th CSCE Intl. Constr. Spec. Conf., CSCE, Montreal, Canada*.

Zhang, R., and El-Gohary, N. (2020). "A machine-learning approach for semantically-enriched building-code sentence generation for automatic semantic analysis." *Proc., Construction Research Congress (CRC) Conf*., ASCE, Reston, VA, 1261-1270.

Zhou, C., He, J., Ma, X., Berg-Kirkpatrick, T., and Neubig, G. (2022). "Prompt consistency for zero-shot task generalization." arXiv preprint arXiv:2205.00049.

Zhou, Z., Zhang, J., Gong, C., and Wu, W. (2023)." Automatic tunnel lining crack detection via deep learning with generative adversarial network-based data augmentation." *Underground Space*, *9*, 140-154.