Korean Journal of Construction Engineering and Management
Vol. 26, No. 4, pp. 089-098 / July 31, 2025
https://dx.doi.org/10.6106/KJCEM.2025.26.4.089

pISSN 2005-6095
eISSN 2465-9703
www.kjcem.org

Predicting Construction Schedule Delays Using Large Language Models

Saruul Ishdorj[1] · Jongho Lee[2] · Young hun Jun[3] · Yongseok Choi[4] · Kyu Heong Kim[5] · Junghon Jeon[6] · Jaeyoon Kim[7] · Jinwoo Kim[8*]

[1]Assistant Professor, Department of Architectural Engineering, Gachon University
[2]Team Leader, Process Management Department, Jehovah Corporation
[3]Senior Researcher, Department of Construction Policy Research, Korea Institute of Civil Engineering and Building Technology
[4]General Manager, Process Management Department, Jehovah Corporation
[5]Chief Executive Officer, Jehovah Corporation
[6]Assistant Professor, Department of Architectural Engineering, Pusan National University
[7]Assistant Professor, Department of Civil Engineering, The University of Texas at El Paso
[8]Assistant Professor, Department of Architectural Engineering, Kumoh National Institute of Technology

Abstract : This study investigates the performance of Large Language Models (LLM), specifically generative pretrained transformer, in predicting construction schedule delays through in-context learning approaches. Predictions were conducted both on unstructured data and on data transformed into structured formats, allowing for a comparative analysis of prediction accuracy across different levels of data preprocessing. The results showed that the prediction accuracy of the zero-shot approach was limited to 38.3%. In contrast, the few-shot approach achieved a prediction accuracy of up to 48.8% when trained on unstructured data and up to 73.9% when trained on structured data. These findings demonstrate the critical role of data structuring and prompt design in enhancing model performance. This study contributes to advancing the use of LLM in construction delay analysis and offers practical implications for improving schedule management and decision-making processes in the construction industry.

Keywords : Large Language Models, Construction Delay, Generative Pre-trained Transformer, Prediction Models, Prompt Engineering

1. Introduction

1.1 Research Background

Construction projects are inherently vulnerable to schedule delays due to a wide range of factors, often resulting in serious consequences such as increased costs, quality degradation, and stakeholder conflicts. As a result, the early identification and accurate prediction of schedule delays are regarded as critical tasks in construction management. Although numerous studies have employed statistical models, simulation-based techniques, and machine learning approaches to address this issue, delays continue to occur frequently in practice (Lee et al., 2018). In fact, 85% of large-scale construction projects in South Korea failed to meet their scheduled deadlines, and only 15% were completed on time between 2005 and 2021 (Lee et al., 2024). To overcome these limitations, this study proposes a predictive model for estimating construction delay durations using real-world data from domestic projects and a Generative Pre-trained Transformer (GPT) based Large Language Models (LLM).

LLM have proven effective across various industries, the application of generative AI to address domain specific problems in construction such as schedule delays remains in its early stages (Smetana et al., 2024). Therefore, this study aims to evaluate the predictive accuracy of GPT in estimating delay durations based on its ability to analyze and interpret textual data.

* **Corresponding author**: Jinwoo Kim, Department of Architectural Engineering, Kumoh National Institute of Technology, Global Building G457, 61 Daehak-ro, Gumi-si, Gyeongsangbuk-do, 39177, Korea
E-mail: jinwoo@kumoh.ac.kr
Received December 24, 2024; **revised** April 7, 2025
accepted June 16, 2025

Specifically, it preprocesses unstructured delay records into structured formats suitable for model input and compares the prediction accuracy of the GPT model across different data types. The findings of this research are expected to contribute to the development of a systematic and effective tool for the early prediction and proactive management of construction schedule delays.

1.2 Framework and Scope

As mentioned above, this study systematically investigates the prediction of construction delay durations using LLM such as GPT. It also analyzes differences in model performance based on data types (i.e., structured vs. unstructured) and learning methods (i.e., zero-shot vs. few-shot). The research procedure and scope are outlined as follows.

First, real-world construction delay data were collected from public construction projects in South Korea. The dataset consists of 150 projects that were either completed or ongoing over the past 20 years. Due to the unstructured nature of the collected data, additional preprocessing was necessary. To address this, expert input was used to hierarchically classify the causes of delays, and analytic hierarchy process (AHP) techniques were employed to assign weights to each contributing factor. This process transformed the unstructured textual data into a structured format, enabling the construction of a dataset suitable for training LLM.

Finally, to predict delay durations, the study applied in-context learning methods using both zero-shot and few-shot approaches with the GPT model. Model performance was evaluated by comparing prediction accuracies under each method. Specifically, both raw unstructured and structured data formats were used to analyze how model performance varied depending on prompt design comparing simple baseline prompts with enhanced prompt structures.

2. Literature review

2.1 Previous Studies on Construction Schedule Delays

Schedule delays remain a persistent challenge in the construction industry, prompting various studies to analyze delay and risk factors in order to find effective solutions. For isntance, Larsen et al. (2016) analyzed public construction projects in Denmark and identified major causes of delays, including uncertainty in financial management, administrative inefficiencies, inadequate planning, and execution errors during construction.

Cho et al. (2021) investigated 19 precast concrete projects in South Korea, examining the relationship between initial project conditions and schedule delays. Similarly, Kayelle et al. (2023) applied an AHP to evaluate and prioritize delay factors in Brazilian public road infrastructure projects. Among early studies incorporating generative AI, Ghimire et al. (2024) suggested that generative models can support document management, automated design, and schedule optimization, while also highlighting data quality and ethical concerns as unresolved challenges. In a related effort, Baek et al. (2023) developed a GPT based framework that automatically detects conflict-inducing factors in the civil engineering and construction management domains through the analysis of news articles.

However, many of these existing studies tend to focus on specific project types or construction methods, thereby limiting the generalizability and applicability of their findings across broader project contexts.

2.2 Current Applications of Generative AI in the Construction Industry

Recently, the use of generative AI technologies in the construction industry has seen rapid growth. Ploennigs and Berger (2024) evaluated the effectiveness of AI-based image generators such as GPT, DALL·E, and MidJourney in enhancing the performance of Building Information Modeling (BIM) and computational design workflows. Their study demonstrated that generative AI can be applied not only to visual content generation but also to the broader architectural design process.

Aalaei et al. (2023) explored the application of AI and deep learning techniques to improve efficiency during the early design stages. Similarly, Qin et al. (2024) proposed a new design and optimization framework using LLM to address recurring issues related to quality and usability in traditional design approaches. He et al.

(2024) further advanced this line of work by introducing Generative ABIM, a framework that integrates AI and BIM to overcome limitations in existing AI-based design methods.

Table 1. Results of previous studies on generative AI in the construction industry

Year	Authors	Results and Performance
2024	He et al.	Experimental results of the proposed PCDM demonstrated substantially superior performance compared to the latest StructGAN.
2024	Nyqvist et al.	Generative AI solutions were applied in project management, risk mitigation, and decision-making processes, achieving tangible performance improvements.
2024	Parsafard et al.	Accuracy results: RoBERTa - 91%, GPT-3.5 - 89%, BERT - 87%.
2024	Ploennigs & Berger	Improved the validity of generated floor plans by up to 90%.
2024	Qin et al.	Design speed was 10 times faster than human engineers, and while AI designs showed an average material consumption difference of 10-30%, they were deemed suitable within realistic ranges.
2024	Ray et al.	Analysis of OSHA accident reports showed that 69% of cases matched human analysis exactly, resolving the potential for hallucination effects,
2024	Yoo et al.	A fine-tuned GPT model achieved an impressive overall accuracy of 82.33% in predicting construction accidents, forming a clearly superior model compared to existing AI approaches.
2023	Aalaei et al.	Performed simulations on the RPLAN dataset with high accuracy, establishing itself as a useful tool for generating architectural floor plans,
2023	Prieto et al.	Overall performance was adequate, though somewhat limited in tasks such as cost estimation.

In the area of construction management, Ray et al. (2024) analyzed whether generative AI could effectively replicate human-level analysis when interpreting construction accident reports. Parsafard et al. (2024) examined the potential of generative AI in automating quantity take-offs and enhancing cost management processes. Prieto et al. (2023) evaluated the feasibility and limitations of applying generative AI to the automated generation of construction schedules. Regarding safety-focused research, Yoo et al. (2024) developed a GPT-based safety prediction model using transfer learning to classify six types of construction accidents. Most recently, Nyqvist et al. (2024) investigated the integration of generative AI solutions into broader construction project management systems.

The findings and performance metrics of these studies are summarized in 〈Table 1〉.

These prior studies collectively demonstrate that generative AI technologies can play significant roles across a wide range of domains in construction, including design, data analysis, and project management. However, their application to the specific problem of schedule delay prediction remains underexplored, highlighting the need for further research in this area.

3. Research Methodology

3.1 Data Collection and Characteristics

3.1.1 Data Collection

In South Korea, it is generally difficult to collect comprehensive data on construction schedule delays due to limitations in public data disclosure and the absence of a nationwide standardized data management system. Most data are managed only in certain major cities, and schedule delays are often perceived negatively in the construction industry, which further hinders open data sharing. As a result, obtaining consistent nationwide data sets poses significant challenges.

Given these constraints, this study focused on publicly accessible data and collected schedule extension records from 150 public construction projects that were either completed or ongoing over the past 20 years. These records included approximately 309 instances of formal delay modifications.

After excluding 12 cases with missing key information, a total of 297 valid cases were selected for analysis. These are summarized in 〈Table 2〉. The final dataset encompasses a wide variety of project types, including building projects such as apartment complexes and general commercial facilities, as well as civil engineering works such as roads and bridges. This allowed the study to identify both common causes of delays and project-type-specific characteristics in a systematic manner. Moreover, the diversity and comprehensiveness of the dataset significantly contributed to the reliability of the delay prediction model developed and analyzed in this study.

Table 2. Construction delay cases

Phases of Delay Occurrence	Project types		
	Building	Civil engineering	Total cases
Pre-construction	6	10	16
Mid-construction	158	114	272
Post completion	1	8	9
Total cases	165	132	297

3.1.2 Characteristics and Limitations of the Data

As illustrated in ⟨Fig. 1⟩, each delay case record consists of multiple fields, including project name, construction type, project budget, revision sequence, number of extension days, and the stated reasons for the delay. However, one of the major limitations of the dataset lies in the representation of delay reasons: multiple delay factors are often embedded within a single record of extended duration, and similar causes are inconsistently described across different projects.

This results in a high degree of unstructuredness in the delay reason field, which poses a significant challenge for data-driven modeling. While core project attributes such as construction type, start and end dates, and budget are managed in a typical row and column format and constitute structured data, the delay reason records, written in free text format without a defined schema, are classified as unstructured data.

To address this, additional preprocessing was performed based on expert consultation. The free text delay reasons were systematically categorized and standardized through expert interviews, transforming them into structured labels. Following this categorization, an AHP was applied to derive relative weights for each delay factor.

This preprocessing made it possible to prepare both structured and unstructured datasets for model training. Accordingly, the study proceeded to evaluate and compare the predictive performance of a GPT based LLM trained separately on each data format. This comparative analysis serves to assess the usability of both data types and identify how data structure influences the accuracy of delay duration prediction.

3.2 Data Preprocessing

3.2.1 Data Structuring Process

Because the risk level implied by each stated delay reason differs, this study adopted the AHP as a method to categorize and prioritize merged delay factors based on expert judgment ⟨Fig. 1⟩. Prior to the analysis, potential factors contributing to schedule delays were preliminarily identified through case reviews, literature surveys, and an examination of previous studies. To ensure comprehensive and systematic classification, interviews were conducted with five construction scheduling

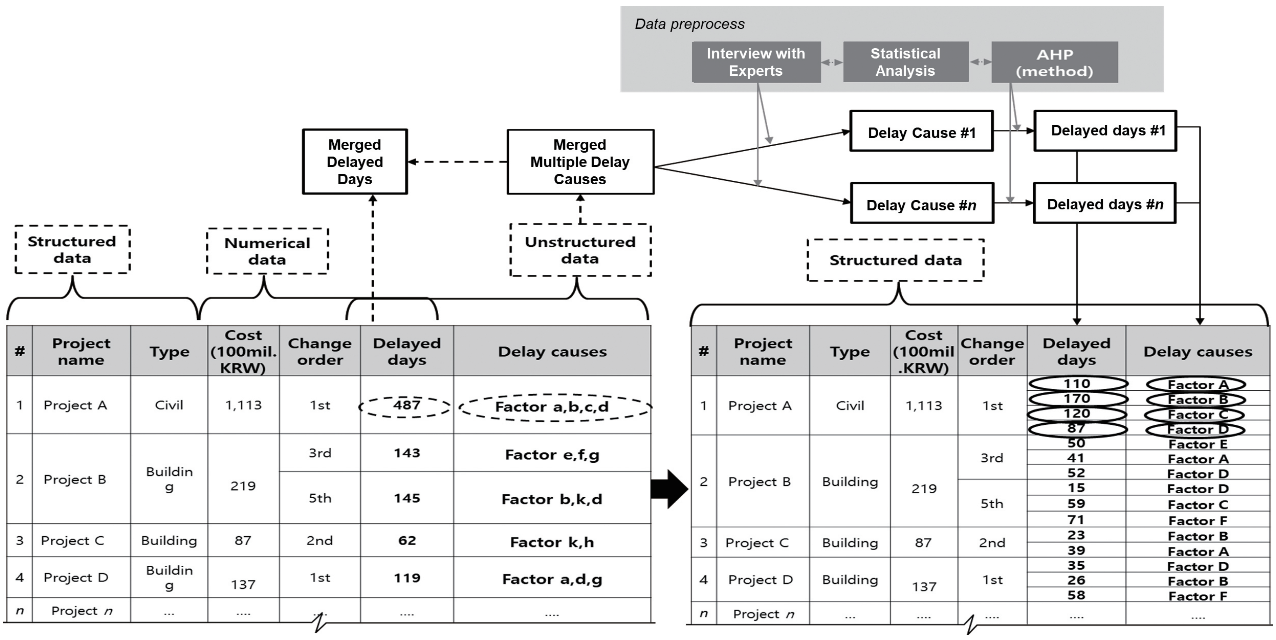

Fig. 1. Data preprocessing framework

experts, based on which the final set of delay categories was established.

For instance, in Project A, the delay reason was originally documented in an unstructured format as follows:

"The schedule was extended due to supplementary design work related to civil complaints and insufficient time to complete the main works."

Such free form descriptions posed challenges for analysis. Accordingly, the delay reason was restructured into two standardized categories: "delay due to civil complaints" and "delay due to insufficient time," making the data analyzable in a structured format. Using these structured categories, AHP was performed to quantify the influence of each delay factor by assigning normalized weights.

AHP surveys were administered to 15 experts. Although prior studies (e.g., Lee, 2002) have noted that meaningful results can be obtained with fewer than 10 experts if the group is homogeneous and possesses substantial professional knowledge in construction schedule management, this study sought to ensure greater objectivity and validity.

Accordingly, the survey was administered to 15 professionals with over five years of relevant experience in schedule estimation and construction project management. The demographic characteristics of the participants are summarized in ⟨Table 3⟩.

Table 3. Demographic characteristics

Characteristic	No. of experts	Frequency (%)
Age		
29~30	5	33
31~40	9	60
41 and over	1	7
Experience		
5~10years	4	27
10~15years	10	67
More than 15years	1	7
Major scope of work		
Architecture	11	73
Civil Engineering	3	20
Equipment	1	7
Major project involvements		
Apartment	2	13
Public Institution Buildings	3	20
Commercial Buildings	3	20
Plant/Factories	7	47

Through the survey, pairwise comparisons were conducted to derive the relative influence of each factor. The resulting AHP derived weights were then applied to the decomposed delay reasons originally embedded within single extension entries, thereby enabling a more logical and structured distribution of delay durations across categorized factors.

Although it is not feasible to perfectly match the adjusted delay days with actual records, the use of derived factor weights contributed to enhanced reliability and consistency from a data analysis perspective.

3.2.2 Framework of Construction Delay Factors

As mentioned earlier, the 297 collected cases of construction schedule extensions were initially grouped into seven major categories based on expert input. These categories served as the foundational criteria for dividing the dataset into training and testing sets. To further refine the classification, delay causes were subdivided from a root-cause perspective into internal and external factors. Internal factors refer to causes originating from stakeholders or conditions within the project itself, whereas external factors stem from entities or environments outside the immediate project scope.

At the most detailed level, individual causal elements responsible for the delay were assigned. Through this hierarchical approach, the dataset was ultimately structured into 16 primary categories and 55 subcategories, enabling more granular analysis of delay types and their sources.

Given that different delay factors may have varying levels of impact on project schedules, an AHP analysis was conducted to quantify their relative influence. Expert participants were asked to perform pairwise comparisons of the identified delay factors, from which relative weights were calculated. These weights were assessed using the Consistency Index (CI), and only results with a CI below 0.05 were retained to ensure reliability and logical coherence of the comparison matrix.

The resulting weights were then used to allocate specific delay durations to individual causes within each project. For instance, if a project was reported to have a 100day delay with a generalized reason such as "design change and material procurement delay," the AHP derived weights were used to apportion the delay: 54

days attributed to design change and 46 days to material procurement delay. This procedure significantly enhanced both the structure and analytical reliability of the delay data, which was essential for preparing the dataset for LLM based predictive modeling.

3.3 Prediction Experiments and Analytic Methods

3.3.1 Data Classification

To facilitate delay prediction, the dataset was further categorized according to the seven major delay factors identified during the expert-driven preprocessing stage. These categories are summarized in 〈Table 4〉. During this classification process, 17 cases with delay reasons that were either ambiguous or not clearly defined—even after expert review—were consolidated under an “Others” category and excluded from the performance evaluation.

By organizing the dataset according to these seven distinct delay factor groups, the experimental design allows for independent assessment of the model’s predictive performance for each category. This structure enhances the interpretability of the results and supports a deeper understanding of how the model performs across various types of delay causes.

Table 4. Number of data divisions by major delay factors

No.	Delay factor	No. of data
1	Occurence of construction suspension	74
2	Design document revision	91
3	Increase in non-working days	40
4	Delay in securing construction site	16
5	Delay in material procurement	24
6	Issues with ensuring adequate construction period	24
7	Extension of completion settlement period	11
8	Others	17

3.3.2 Training and Test Set Partitioning

Each dataset was associated with a specific primary delay factor, allowing the data to be systematically organized in a way that facilitates both analysis and model training. Within each delay factor group, the data were divided into a training set (70%) and a test set (30%).

This stratified division ensured that each test set retained the characteristics of the full dataset, allowing the evaluation to be representative and reducing the risk of bias. By maintaining this ratio consistently across all seven delay factor categories, the model was trained and tested under balanced and realistic conditions, supporting the reliability of the performance comparisons among experimental groups.

3.3.3 Prompt Designs

As shown in 〈Fig. 2〉, prompt engineering plays a critical role in guiding the performance of LLM, as the output quality is highly influenced by the format and structure of the prompt (Yoo et al., 2024). Among various techniques, one widely adopted method is the use of prompt templates, which provide structured guidance to the model through the inclusion of specific elements and formats.

```python
# Define variables for the project details
project_type = "건축"  # Example: '건축' for project type
budget = 50  # Example: 50 billion KRW
delay_factors = ["설계도서(공법) 변경", "공사 중지 발생", "비작업일수 증가"]  # List of delay factors

# Generate the prompt
prompt = f"""
이 프로젝트는 공사 종류가 '{project_type}', 사업비가 '{budget}억원'이며, 주요 지연 요인은 {delay_factors}입니다.
이 프로젝트가 며칠 동안 지연될지 예측해 주세요.
"""

# Print the prompt
print(prompt)
```

Fig. 2. Example of prompt design structure

A zero-shot prompt presents the model with a task without offering any examples. In contrast, an n-shot prompt (i.e., in this study, few-shot) provides one or more input–output examples to help the model infer patterns and generate responses accordingly. Structured prompting, as discussed by Parsafard et al. (2024), involves designing templates based on consistent patterns or formats that guide the model’s interpretation and reasoning.

This study compared the performance of zero-shot and few-shot prompting strategies. In the zero-shot setting, the model was asked to predict the delay duration based solely on the test input, without any reference examples. This approach is suitable for scenarios where prior contextual data are unavailable or limited.

In the few-shot setting, the model was provided with multiple examples representing typical input–output pairs. The prompts were designed to help the model understand the structure of the data and infer delay durations based on learned patterns. Separate prompts were constructed for structured and unstructured data, enabling independent evaluation of each input type.

Each prompt included key project attributes such as project type, budget, and delay cause(s) and asked the model to estimate the number of delay days. These prompts were formulated to simulate real-world conditions while maintaining consistency across input types.

3.3.4 Model Accuracy Evaluation

In this study, the accuracy of the GPT model was evaluated to assess its predictive performance. Accuracy is an indicator that expresses how closely the predicted value approximates the actual value, represented as a percentage (%), and was calculated using the following formula.

$$Accuracy = (1 - \frac{|A_i - P_i|}{A_i}) \times 100 \tag{1}$$

A_1 : $actualvalue$
P_1 : $predictedvalue$

This formula calculates the prediction accuracy of the model based on the difference between the predicted and actual values. The accuracy was computed individually for each data point, and the average of these values was used to evaluate the overall model performance.

4. Results and Discussion

4.1 Results

This study utilized the GPT 4o model to predict construction delay durations and compared the predictive performance between zero-shot and few-shot learning strategies. The experiments were conducted based on zero-shot learning and few-shot learning using both raw unstructured data and structured data, and the differences in performance under two types of prompt designs within each learning strategy were also analyzed. The results are presented in 〈Table 5〉.

1) Zero-shot Learning Results

In this approach, no training data were provided in advance. The model was given only the project type and budget as input, and was asked to estimate the number of delay days based on a given delay cause. As a result, the average prediction accuracy of the zero-shot method was 38.3%, indicating that the model's performance is limited when relying solely on general contextual understanding without prior training.

2) Few-shot Learning Results

After providing the training data, the model was given test inputs with the same questions used in the zero-shot setting. The performance varied depending on the type of training data (i.e., unstructured vs. structured) and the complexity of the prompt design.

When trained on unstructured data, a simple prompt that included only the project type, delay reason, and delay factor resulted in an average prediction accuracy of 41.3%, representing a slight improvement over the zero-shot setting. When the prompt was enhanced to include additional information such as project scale and budget, the average accuracy increased to 48.8%, confirming

Table 5. Prediction results of construction delay days

Delay factor	Actual delay (days)	Zero-shot		Few-shot (unstructured data)				Few-shot (structured data)				
		Prompt design 2		Prompt design 1		Prompt design 2		Actual delay (days)	Prompt design 1		Prompt design 2	
		Pre. Days	Accuracy %	Pre. Days	Accuracy %	Pre. Days	Accuracy %		Pre. Days	Accuracy %	Pre. Days	Accuracy %
Occurence of construction suspension	2691.1	1430	53.1	7127	-64.8	2185	81.2	3495	3500	99.9	3590	97.3
Design document revision	3661.6	795	21.7	1615	44.1	2002	54.7	4831	3210	66.4	3350	69.3
Increase in non-working days	723.1	235	32.5	245	33.9	258	35.7	970	920	94.8	870	89.7
Delay in securing construction site	1410	175	12.4	740	52.5	745	52.8	1410	743	52.7	720	51.1
Delay in material procurement	120.7	150	75.8	225	13.6	270	23.6	192	265	62	290	49
Issues with ensuring adequate construction period	564.5	140	24.8	125	22.1	170	30.1	602	720	80.4	750	75.4
Extension of completion settlement period	145.7	70	48.1	85	58.4	92	63.2	151	210	60.9	230	47.7
Average value			**38.3**		**41.3**		**48.8**			**73.9**		**68.5**

Note: Pre.=Predicted

that more informative prompt design contributed to improved predictive performance under unstructured data conditions.

In contrast, training on the structured version of the data led to significantly higher accuracy. With a simple prompt, the model achieved an average accuracy of 73.9%, indicating that structured data substantially enhanced prediction performance compared to unstructured data. However, when project scale and budget were added to the structured prompt, the average accuracy decreased slightly to 68.5%. This suggests that excessive prompt complexity may reduce the model's predictive efficiency when trained on already structured data.

4.2 Discussion

This study analyzed the impact of LLM on the prediction of construction delay durations, focusing on differences arising from data type (structured vs. unstructured) and learning strategy (zero-shot vs. few-shot). The experimental results yield several meaningful implications.

First, the relatively low accuracy of 38.3% in the zero-shot setting suggests that, without prior exposure to relevant training data, the model struggles to establish a meaningful relationship between delay causes and predicted delay durations. While LLM demonstrate strong general contextual understanding, these findings indicate that domain-specific tasks require tailored learning. Zero-shot prompting may serve as a useful starting point in situations where training data are limited, but its ability to produce accurate predictions remains inherently constrained.

In contrast, under the few-shot setting, the type of training data significantly influenced prediction performance. When trained on structured data, the model achieved an accuracy as high as 73.9%, whereas training on unstructured data yielded a lower maximum of 48.8%. This clearly highlights the crucial role of structured input in providing a logical and interpretable framework that LLM can effectively learn from.

Moreover, the prompt design also had differential effects depending on the data format. For unstructured data, enhanced prompts that included additional project details improved accuracy from 41.3% to 48.8%. However, for structured data, the same enhancement resulted in a slight performance decrease from 73.9% to 68.5%. This suggests that, while detailed guidance can help the model better understand unstructured inputs, overly complex prompts may introduce noise or redundancy when the input is already structured.

These findings emphasize the importance of selecting an optimal prompt strategy based on the nature of the data. This study provides empirical evidence that both data structuring and prompt engineering play a decisive role in determining the predictive performance of LLM (i.e., GPT model).

In particular, the positive impact of data structuring highlights the value of thorough preprocessing in construction project data management and model training pipelines. These results not only contribute to improving delay prediction in construction but also suggest broader applicability of LLM in similar domain-specific forecasting tasks. From a practical standpoint, the study demonstrates that clear and concise prompt design is a key factor in optimizing model performance.

5. Conclusion

This study employed a LLM to predict construction delay durations and analyzed performance differences based on data type (i.e., structured vs. unstructured) and learning strategy (i.e., zero-shot and few-shot). Through this process, the study empirically examined how data structuring and prompt design affect the predictive accuracy of LLM.

The results showed that the zero-shot approach achieved a baseline level of performance without prior training data, but its accuracy was lower compared to learning-based methods. In the few-shot setting, data structuring played a critical role in improving prediction accuracy; models trained on structured data consistently outperformed those trained on unstructured data. Furthermore, detailed prompt design enhanced performance in the unstructured data setting, whereas in the structured data setting, simpler prompts proved to be more effective.

These findings demonstrate that both data structuring and prompt design have a significant impact on the performance of LLM-based predictive models. In particular, structured data serves as a key enabler for GPT models to effectively learn and utilize organized information. This suggests that, beyond construction delay prediction, the potential of LLM can be further unlocked in other domains through appropriate data structuring and optimal prompt engineering.

Overall, the study shows that data preprocessing and LLM application strategies can make meaningful contributions to the development of predictive models for construction delays. Practically, the findings may be applied to improve construction project data management and schedule forecasting. Academically, the results offer insights into how LLM can address domain-specific problems through tailored data preparation and learning approaches.

While this study successfully demonstrated the use of GPT for predicting construction delay durations, several limitations should be acknowledged.

First, the dataset was limited to 297 public construction projects in South Korea, which may not adequately represent the diversity of project types or case volumes. To address this limitation, delay durations were distributed across delay causes using expert derived AHP weights. However, these fixed weights may not fully capture project specific or time sensitive variability, potentially leading to over- or underestimation of certain factors. Future research should aim to build larger datasets that include private sector and international projects, and consider training the weighting system on project level records rather than relying on static AHP values. Applying multitask LLM could also enable the joint prediction of not only delay durations but also the occurrence and frequency of each causal factor.

Second, this study assumed that project scale (i.e., budget) is closely related to schedule delays and to assess its impact on model performance. However, the dataset used in this study included only projects that had experienced actual delays, and the availability of project attribute information was limited. As a result, characteristics such as construction type and specific delay reasons could not be compared in a fully balanced manner.

Third, the evaluation of the GPT model was confined to zero-shot and few-shot learning settings. Fine-tuning and comparative analysis with other predictive models were beyond the scope of this study and are suggested for future work. Nonetheless, this research has demonstrated the potential of GPT and other LLM to play a significant role in construction delay prediction, both practically and academically.

Acknowledgements

This research was supported in part by the 2024 Research and Development Reserve Fund Program (Academic-Industrial Collaboration Project) of the Korea Institute of Civil Engineering and Building Technology (KICT), under the project titled "Development of Risk Diagnosis, Prediction, and Response Solutions for Construction Schedule Delays Using Large Language Models" (Project No. 20240385-001, Program Type: 24-Reserve-Institution-Academic-Type 1).

The authors would also like to thank Changgeon Lee, Undergraduate Research Assistant at the Kumoh National Institute of Technology Construction Engineering & Management Lab (KITCEM), for his assistance with English translation.

ORCID

Saruul Ishdorj, https://orcid.org/0000-0002-4455-174X
JungHo Jeon, https://orcid.org/0000-0001-6816-0092
Jaeyoon Kim, https://orcid.org/0000-0002-0723-0664
Jinwoo Kim, https://orcid.org/0000-0001-7003-1222

References

Aalaei, M., Saadi, M., Rahbar, M., and Ekhlassi, A. (2023). “Architectural layout generation using a graph-constrained conditional generative adversarial network (GAN).” *Automation in Construction*, 155, 105053. https://doi.org/10.1016/j.autcon.2023.105053

Baek, S., Namgoong, D., Won, J., and Han, S.H. (2023). “Automated Detection of Social Conflict Drivers in Civil Infrastructure Projects Using Natural Language Processing.” *Applied Sciences*, 13(20), 11171. https://doi.org/10.3390/app132011171

Cho, K., Ahn, S., Park, K., and Kim, T.W. (2021). "Schedule delay leading indicators in precast concrete construction projects: Qualitative comparative analysis of Korean cases." *Journal of Management in Engineering*, 37(4), 04021024. https://doi.org/10.1061/(ASCE)ME.1943-5479.0000915

Ghimire, P., Kim, K., and Acharya, M. (2024). "Opportunities and challenges of generative ai in construction industry: Focusing on adoption of text-based models." *Buildings*, 14(1), 220. https://doi.org/10.3390/buildings14010220

Kayelle, E., Rabbani, E.K., and Macedo, M. (2023). "Identifying the Causes of Delay Using the AHP (AHP) Method in Brazilian Public Road Infrastructure Project." *J. Mgmt. & Sustainability*, 13(2), 45. https://doi.org/10.5539/jms.v13n2p45

Larsen, J.K., Shen, G.Q., Lindhard, S.M., and Brunoe, T.D. (2016). "Factors affecting schedule delay, cost overrun, and quality level in public construction projects." *Journal of Management in Engineering*, 32(1), 04015032. https://doi.org/10.1061/(ASCE)ME.1943-5479.000039

Lee, D.H., Jeong, I.S., Park, J.W., and Kim, K.H. (2018). Development of Key Performance Index (KPI) and Business Model for Construction Projects, KICT Research Report, 2018-12

Lee, K., Ding, Z., Im, J.-B., Park, S., and Kim, J.-H. (2024). "Investigation of construction duration influence factors and comparison of AI methods for estimating appropriate construction duration in public sewerage construction." *Korean Journal of Construction Engineering and Management*, KJCEM, 25(6), 066-074. https://doi.org/10.6106/KJCEM.2024.25.6.066

Nyqvist, R., Peltokorpi, A., and Seppänen, O. (2024). "Integration of generative AI across construction management." *IOP Conference Series: Earth and Environmental Science*, 1389, 012011. https://doi.org/10.1088/1755-1315/1389/1/012011

Parsafard, P., Elezaj, O., Ekundayo, D., Vakaj, E., Parmar, M., and Wani, M.A. (2024). "Automation in Construction Cost Budgeting using Generative AI." *Proceedings of the International Conference on Industrial Engineering and Operations Management*, IEOM Society International, 14, 1713-1722. https://doi.org/10.46254/AN14.20240466

Ploennigs, J., and Berger, M. (2024). "Automating computational design with generative AI." *Civil Engineering Design*, 6(2), 41-52. https://doi.org/10.48550/arXiv.2307.02511

Prieto, S.A., Mengiste, E.T., and García de Soto, B. (2023). "Investigating the use of ChatGPT for the scheduling of construction projects." *Buildings*, 13(4), 857. https://doi.org/10.3390/buildings13040857

Qin, S., Guan, H., Liao, W., Gu, Y., Zheng, Z., Xue, H., and Lu, X. (2024). "Intelligent design and optimization system for shear wall structures based on large language models and generative AI." *Journal of Building Engineering*, 95, 109996. https://doi.org/10.1016/j.jobe.2024.109996

Ray, U., Arteaga, C., and Park, J. (2024). "Comparative analysis of cognitive agreement between human analysts and generative AI in construction safety risk assessment." *Proceedings of the International Symposium on Automation and Robotics in Construction*, ISARC, 41, 452-458. https://doi.org/10.22260/ISARC2024/0059

Smetana, M., Salles de Salles, L., Sukharev, I., and Khazanovich, L. (2024). "Highway Construction Safety Analysis Using Large Language Models." *Applied Sciences*, 14(4), 1352.

Uddin, S. J., Albert, A., Ovid, A., and Alsharef, A. (2023). "Leveraging ChatGPT to aid construction hazard recognition and support safety education and training." *Sustainability*, 15(9), 7121. https://doi.org/10.3390/app14041352

Yoo, B., Kim, J., Park, S., Ahn, C.R., and Oh, T. (2024). "Harnessing Generative Pre-Trained Transformers for Construction Accident Prediction with Saliency Visualization." *Applied Sciences*, 14(2), 664. https://doi.org/10.3390/app14020664