How well do large language models perform in green building assessment? An evaluation analysis from a case data set

Qiufeng He[1,2,3], Zezhou Wu[1,2,3*] and Xiangsheng Chen[1,2,3]

[1]State Key Laboratory of Intelligent Geotechnics and Tunnelling, Shenzhen, China.
[2]Key Laboratory for Resilient Infrastructures of Coastal Cities (Shenzhen University), Ministry of Education, Shenzhen, China.
[3]Sino-Australia Joint Research Center in BIM and Smart Construction, Shenzhen University, Shenzhen, China
Corresponding author's E-mail: wuzezhou@szu.edu.cn

Abstract

Recent advances in Large Language Models (LLMs) have demonstrated their impressive capabilities in various tasks. However, their potential in the specialized field of green building assessment has not been explored. Such a study is necessary to understand their performance in this domain, with the goal of optimizing LLMs to reduce the workload of manual green building assessments and enable designers to conduct preliminary self-assessments more efficiently and economically. In this regard, this study expands the dataset from 112 to 1200 real-world cases, and then eleven leading LLMs are selected for evaluation using both long and short text inputs combined with three different prompt engineering techniques (i.e., zer0-shot, zero-shot CoT, few-shot) to determine their accuracy. The findings indicate that LLMs perform better with short text inputs, particularly GPT-4, which showed the highest effectiveness in the green building evaluation field. Prompt engineering improved the performance of GPT-4 with short text inputs, though its effectiveness varied across different LLMs. Furthermore, LLMs excelled in evaluating qualitative criteria that do not require logical reasoning but performed poorly in assessing quantitative criteria that involve complex mathematical calculations. Research findings provide valuable insights for future development of LLM-based methods for green building evaluation, aiming to alleviate current manual assessment burdens and improve design review processes.

Keywords: Large language models, green buildings, green score, automatic evaluation, prompt engineering.

1. INTRODUCTION

The construction industry significantly contributes to carbon emissions and resource wastage (He, Wu, Wu, et al., 2024), accounting for 37% of global energy use and greenhouse gas emissions (Liu et al., 202). Green buildings mitigate these impacts by enhancing energy efficiency, conserving resources, and promoting occupant well-being (He et al., 2023). Consequently, many countries, particularly China, now mandate that new constructions meet green building evaluation standards (Olanrewaju et al., 2024). These standards, tailored to specific environmental and geographical contexts, present evaluation challenges due to their complexity, necessitating high expertise and time. Automated evaluation tools are needed to address the growing demand and resource limitations.

Previous scholars have explored various methods for automating green building evaluations, including: (1) coding green building evaluation standards to calculate green scores (Dubljević et al., 2023; Chen et al., 2017; Nguyen et al., 2016; Dubljević et al., 2024); (2) predicting green scores using machine learning algorithms (Juan et al., 2022; Jalaei et al., 2020; Ramakrishnan, Liu, et al., 2023); and (3) inferring green scores through rule-based natural language processing (NLP) methods by compiling evaluation standards into ontological knowledge graphs and SWRL rules (He, Wu, and Chen, 2024; Jiang et al., 2018). These methods have limitations. code compilation lacks flexibility and requires updates with changing standards; machine learning requires vast data and may not capture semantic nuances effectively; and rule-based NLP methods rely on predefined knowledge graphs that may not cover all complexities.

Large language models (LLMs) such as ChatGPT, BERT, and ELMo, pretrained on extensive data, offer potential solutions by capturing semantic dependencies in lengthy texts like green building standards and case details. However, applying LLMs directly to green building evaluation can lead to computational errors. Evaluating LLM performance in this context is crucial to understand error rates and improve application reliability.

Similar LLM performance evaluations have been conducted in other fields. For instance, Jahan et al. (2024) assessed GPT-3.5, PaLM-2, and LLaMA-2 on six different biomedical tasks (26 datasets). Shojaee-Mend et al. (2024) compared ChatGPT, Bard, and Claude on answering neurophysiology questions in Persian and English. Ammar et al. (2024) evaluated LLaMA-7b, JAIS-13b, and GPT-3.5-turbo on 10,813 commercial court cases in Arabic legal judgments. However, research on LLMs in the green building assessment is lacking. Such studies can help understand LLMs' effectiveness, providing insights for cost-effective and efficient LLM-based green building assessments. In this regard, this study aims to evaluate LLMs' application performance in green building evaluation. It helps understand LLMs' text processing and evaluation capabilities and limitation.

2. METHOD AND MATERIAL

This study firstly selects six types of mainstream LLMs, totaling 11 models, as shown in Table 1, and constructs a dataset based on actual cases. Next, the LLMs are queried using three different prompt engineering. The models' evaluation scores are compared with expert scores to determine the accuracy of each model in assessing green building performance. Finally, the accuracy results are analyzed.

Table 1. Mainstream large language models

Model series	Large language models
GPT	GPT-3.5, GPT-4
Claude	Claude-3
Gemini	Gemini-Pro
Llama-2	Llama-2-7B, Llama-2-13B, Llama-2-70B
Mistral	Mistral-Large, Mistral-Medium, Mixtral-8x7B-Chat
Qwen	Qwen-72b-Chat

The data used for testing includes 112 cases, containing both textual descriptions and expert scores, are evaluated according to the Green Building Evaluation Standard (GB/T50378-2019). Based on these cases, six industry experts are invited to supplement the dataset. Three experts revised the 112 cases based on evaluation experience and principles of diversity and practical design, expanding them to 1200 cases. The remaining three experts evaluated the revised cases according to the green building standards. This resulted in a dataset of 1200 cases, including design descriptions and scores.

By interpreting the green building standards and actual case descriptions, the data can be classified as follows: (1) Type I. Quantitative evaluation item directly included in the case data; (2) Type II. Quantitative evaluation item indirectly included, requiring data calculations; (3) Type III. Qualitative evaluation item directly described in the cases; (4) Type IV. Qualitative evaluation item indirectly described, needing an understanding of the case's attributes. Table 2 lists examples of evaluation criteria and case descriptions for these four categories.

Table 2. Mainstream large language models

Type	Content of an evaluation item	Case description corresponding to this item
Type I	8.2.2: Annual runoff control rate of the planned site is 55% for 5 points; 70% for 10 points.	The total annual runoff control rate is 85%.
Type II	6.2.5.1: Outdoor fitness area is at least 0.5% of the total land area for 3 points.	The total land area is 30,100 square meters. The outdoor fitness area includes a basketball court (82 square meters) and a tennis court (24 square meters).

Type III	4.2.3.1: Using glass with safety features for 5 points.	The building uses ZLcjj005 laminated glass from Zhongli brand, which is explosion-proof, waterproof, soundproof, and insulated.
Type IV	5.2.4.1: Measures to ensure water storage does not deteriorate for 5 points.	The drinking water tank and tower in the building have independent structures. Drain pipes are placed below the water tanks. The water tanks have locked manholes, with measures to prevent biological entry and disinfection facilities.

3. EVALUATION PIPELINE

In the PyCharm development environment, the LLMs mentioned in Section 2.2 are evaluated by invoking the relevant codebase or application interface through the Python programming language. For the evaluation of the open-source LLM, the text generation process is implemented with full precision on two Tesla A100 Gpus (80GB). For closed-source LLMs, the application interfaces of these models, such as the OpenAI API and the Google API, are invoked to perform text generation tasks in the form of dialogues.

Considering that a LLM may have knowledge forgetting, that is, the case content exceeds the maximum text length that the model can process, and the LLM's memory ability and attention mechanism will be challenged when processing very long text, so the input of case information to the LLM is carried out in two situations, including:

(1) Inputting one case at a time. Input 1200 cases into the LLM in turn, and let the LLM conduct a green assessment of the content of each case according to the green building evaluation standard.

(2) Inputting each standard item at a time. Each case is divided into 115 items according to the evaluation standard, and each item of the case is entered into a LLM to be evaluated.

Through the above two information input methods, the overall understanding ability, detail processing ability and information retention and extraction ability of 11 kinds of LLMs in green building evaluation could be effectively evaluated.

In addition, considering that the quality of the answers in a LLM would depend on the input instructions, three main prompt engineering are adopted in this study, including:

(1) Zero-shot. Green building case data and green building evaluation standards are fed directly into the LLM. The LLM is then asked to provide a total green score or individual green scores for each item based on the evaluation standards. This prompt engineering assesses the LLM's ability to generate results without any additional prompts or guidance.

(2) Zero-shot CoT. Given the complex logical reasoning challenges of the green building scoring process, the zero-shot thought chain prompt is constructed to follow each direct question with the activation command "Let's think step by step" to guide the LLM to deeper logical reasoning and help it better understand and handle complex tasks.

(3) Few-shot. Before each zero-shot question, the LLM is provided with six examples for learning. Each example includes a case description, an evaluation standard, and the actual correct score for that case. The LLM then assesses the total green score or individual green scores for each item.

4. RESULT ANALYSIS

4.1. Evaluation results of long text input and short text input

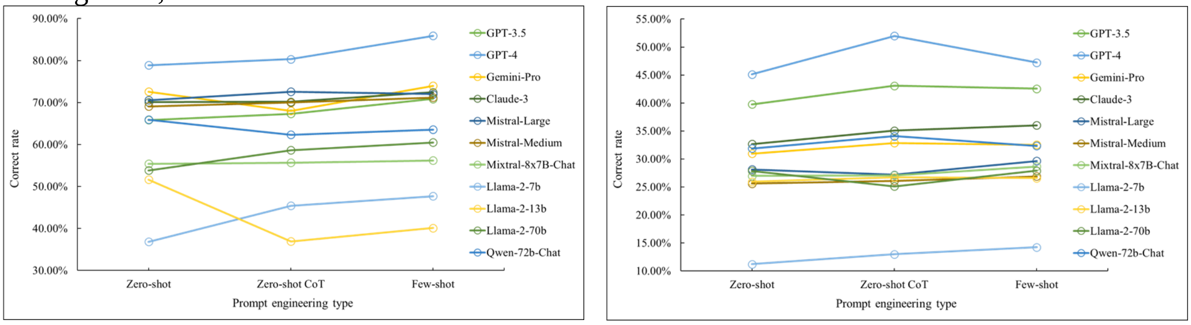
Based on the case input methods from Section 3, the average accuracy rates of 11 LLMs using four prompt engineering types are summarized in Table 3. The findings indicate superior performance of LLMs with short text inputs in green building evaluations. This is primarily due to case information often exceeding thousands of characters, posing limitations on LLMs' memory capacity. Exceeding this limit may lead to information loss, thereby impacting evaluation accuracy. Moreover, in lengthy texts, evaluation details for each item may be widely dispersed, requiring LLMs to extract logical relationships from extensive text and manage multiple information points, potentially resulting in decreased evaluation accuracy. Among mainstream LLMs, GPT-4 demonstrates the highest performance in green building case evaluations, while the Llama-2 series performs least effectively. This disparity is mainly attributed to differences in technical features and capabilities in handling complex tasks. GPT-4, as the latest generation LLM from OpenAI, boasts a larger context window and stronger reasoning abilities, enabling better management of lengthy texts and intricate logical relationships. Leveraging a larger dataset and advanced training methods during pre-training enhances its capacity for precise understanding and evaluation of diverse standards and detailed case specifics. In contrast, the Llama-2 series exhibits comparatively weaker skills in logical reasoning and multi-step processing, posing challenges in accurately assessing complex green building evaluation standards and interwoven case details.

Table 3. Mainstream large language models

LLMs	Evaluation accuracy under long text input	Evaluation accuracy under short text input
GPT-3.5	41.80%	68.01%
GPT-4	53.10%	81.72%
Gemini-Pro	32.12%	71.53%
Claude-3	34.58%	70.90%
Mistral-Large	28.33%	71.73%
Mistral-Medium	26.20%	70.08%
Mixtral-8x7B-Chat	27.57%	55.77%
Llama-2-7b	12.82%	43.32%
Llama-2-13b	26.42%	42.89%
Llama-2-70b	26.97%	57.66%
Qwen-72b-Chat	32.77%	63.94%

4.2. Evaluation results under different prompt engineering

Under short text input, the LLMs' evaluation accuracy with different prompt engineering methods is shown in Figure 1(a). GPT-4 and Gemini-Pro exhibit consistent improvement across all prompt engineering methods, indicating their effective use of contextual information and examples for more accurate reasoning and evaluation. Claude-3 and GPT-3.5 show significant improvement with few-shot but some fluctuation with zero-shot CoT. Under long text input, the evaluation accuracy of LLMs with different prompt engineering methods is shown in Figure 1(b). Similar to short text input, the evaluation results vary with different prompts. Overall, compared to the baseline (i.e., zero-shot), zero-shot CoT and few-shot do not universally improve the accuracy of green building evaluations for all LLMs. Prompt engineering (especially zero-shot CoT and few-shot) requires strong logical reasoning and contextual understanding. Some models lack these capabilities, resulting in limited improvement in evaluation performance. Additionally, with long text inputs, most LLMs perform better with zero-shot CoT, indicating that chain-of-thought guidance is more suitable for evaluating long-text green building cases, whereas few-shot is more effective for short-text evaluations.



(a) Short text input (b) Long text input

Figure 1. Evaluation accuracy of different LLMs under different prompt engineering

4.3. Evaluation results for different standard items

Due to potential significant evaluation errors in long-text case inputs due to the limitations of LLM context window size, this section analyses data under short-text inputs. Following the classification of green building assessments in Section 2.3, averaging the evaluation accuracy of 11 LLMs across three prompt engineering methods for these categories yields the results of LLMs' evaluation accuracy for different items, as shown in Figure 2. It is observed that the LLMs perform best in evaluating case descriptions like Type III, with accuracy rates ranging from 85% to 100%. Evaluation accuracy for case descriptions like Type IV and Type I is moderate, ranging from 50% to 70%, while accuracy for Type II is poorest, ranging from 10% to 40%. Combining the characteristics of each type, it is evident that in green building assessment, LLMs excel in textual content analysis compared to computational tasks, and their application capability weakens with increasing computational complexity. This finding aligns with previous research indicating that LLMs have relatively weaker computational abilities.

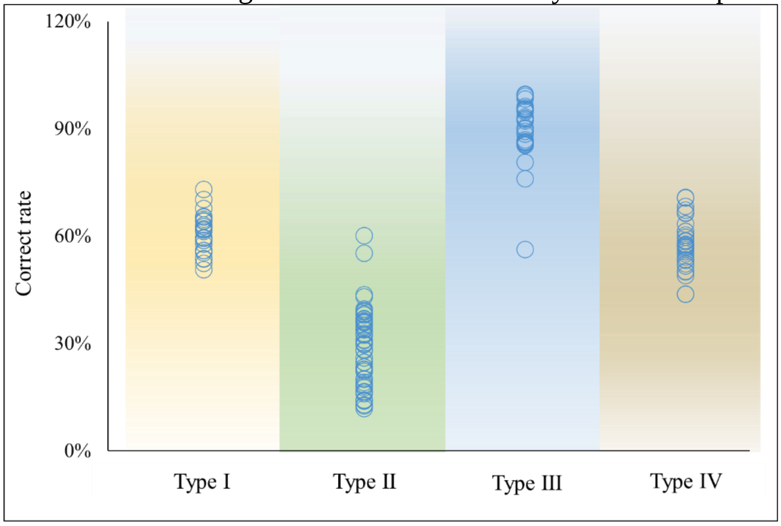


Figure 2. Accuracy of LLMs for evaluating different standard items

5. CONCLUSIONS AND FUTURE RESEARCH

This study evaluates the applicability of current mainstream LLMs in assessing green building performance, drawing the following conclusions:

(1) LLMs exhibit higher error rates when assessing green building performance in long-text cases due to limitations in window length and knowledge retention. Future applications requiring LLMs for assessing green building performance should concentrate information points in case texts and utilize instructions for segmented processing to enhance overall evaluation accuracy.

(2) Among mainstream LLMs, GPT-4 demonstrates the best performance in evaluating green building performance due to its larger context window, stronger reasoning capabilities, and utilization of advanced training methods with extensive datasets during pre-training. Conversely, the Llama-2 series shows the poorest performance in assessing green building performance.

(3) Compared to the baseline (zero-shot), zero-shot CoT and few-shot prompt engineering do not universally enhance the accuracy of green building assessment for all LLMs. Prompt engineering, especially zero-shot CoT and few-shot, necessitates models with strong logical reasoning and contextual understanding abilities. Some models may lack in these aspects, thereby limiting the enhancement of their green assessment performance.

(4) LLMs perform best in assessing qualitative clause items that do not require logical reasoning, while their performance is weakest in assessing quantitative clause items that involve complex mathematical computations. Future strategies could involve instructing LLMs to convert computational processes into code using an embedded Python compiler to reduce their workload during green building assessment.

The evaluation results of this study provide insightful guidance for future development of LLM-based methods for green building assessment, aiming to enhance performance efficiently and economically. Additionally, this study serves as a benchmark for evaluating the application performance of emerging LLMs. Future research directions could explore additional prompt engineering types to determine which are most suitable for LLMs in assessing green building performance. Furthermore, investigating fine-tuning, knowledge enhancement, and constructing thought chains can determine the most effective and cost-efficient methods for improving LLM performance in green building assessment.

6. REFERENCES

Ammar A, Koubaa A, Benjdira B, Nacar O, and Sibaee S. 2024. Prediction of Arabic Legal Rulings Using Large Language Models. Electronics, 13.

Chen P-H, and Nguyen T C. 2017. Integrating web map service and building information modeling for location and transportation analysis in green building certification process. Automation in Construction, 77, pp.52-66.

Dubljević S, Tepavčević B, Markoski B, and Anđelković A S. 2023. Computational BIM tool for automated LEED certification process. Energy and Buildings, 292, pp.113168.

Dubljević S, Tepavčević B, Stefanović A, and Anđelković A S. 2024. BIM to BREEAM: A workflow for automated daylighting assessment of existing buildings. Energy and Buildings, 312: 114208.

He Q, Wu J, Wu Z, Zhang J, and Chen X. 2024. Evolutionary game analysis of prefabricated buildings adoption under carbon emission trading scheme. Building and Environment, 249.

He Q, Wu Z, and Chen X. 2024. An integrated framework for automatic green building evaluation: A case study of China. Frontiers of Engineering Management.

He Q, Wu Z, Li S, Li H, and Wang Y. 2023. Two decades of the evolution of China's green building policy: insights from text mining. Building Research and Information, 51, pp.158-78.

Jahan I, Laskar M T R, Peng C, and Huang J X. 2024. A comprehensive evaluation of large Language models on benchmark biomedical text processing tasks. Computers in Biology and Medicine, 171, pp.108189.

Jalaei F, Jalaei F, and Mohammadi S. 2020. An integrated BIM-LEED application to automate sustainable design assessment framework at the conceptual stage of building projects. Sustainable Cities and Society, 53, pp.101979.

Jiang S, Wang N, and Wu J. 2018. Combining BIM and Ontology to Facilitate Intelligent Green Building Evaluation. Journal of Computing in Civil Engineering, 32.

Juan Y-K, and Lee P-H. 2022. Applying data mining techniques to explore technology adoptions, grades and costs of green building projects. Journal of Building Engineering, 45, pp.103669.

Liu G, Chen R, Xu P, Fu Y, Mao C, and Hong J. 2020. Real-time carbon emission monitoring in prefabricated construction. Automation in Construction, 110, pp.102945.

Nguyen T H, Toroghi Sh H, and Jacobs F. 2016. Automated Green Building Rating System for Building Designs. Journal of Architectural Engineering, 22, pp. A4015001.

Olanrewaju O I, Enegbuma W I, and Donn M. 2024. Operational, embodied and whole life cycle assessment credits in green building certification systems: Desktop analysis and natural language processing approach. Building and Environment, 258, pp.111569.

Ramakrishnan J, Liu T, Zhang F, Seshadri K, Yu R, and Gou Z. 2023. A decision tree-based modeling approach for evaluating the green performance of airport buildings. Environmental Impact Assessment Review, 100, pp.107070.

Shojaee-Mend H, Mohebbati R, Amiri M, and Atarodi A. 2024. Evaluating the strengths and weaknesses of large language models in answering neurophysiology questions. Scientific Reports, 14, pp.10785.

Zhao Y, Liu L, and Yu M. 2023. Comparison and analysis of carbon emissions of traditional, prefabricated, and green material buildings in materialization stage. Journal of Cleaner Production, 406, pp.137152.