Journal of World Architecture, 2026, Volume 10, Issue 4
http://ojs.bbwpublisher.com/index.php/JWA
Online ISSN: 2208-3499
Print ISSN 2208-3480

Evaluating Large Language Models in Green Building Design Decisions: A Case-Based Analysis of Energy, Water, and Material Trade-Offs

Reem AL-qawas, Dan Shao*

School of Art and Design, Dalian Polytechnic University, 116034, Dalian, China

Author to whom correspondence should be addressed.

Abstract: *Background*: AI has shown potential across technical professions, yet its application to green building design—navigating trade-offs between energy efficiency, water conservation, and materials lifecycle assessment (LCA)—remains unevaluated. No prior study has benchmarked frontier large language models (LLMs) using a professor-validated instrument combining multiple-choice question (MCQ) and True/False (T/F) items probing context-dependent sustainability reasoning. *Objective*: To comparatively benchmark four frontier LLMs—Claude, ChatGPT-5.2, DeepSeek-R1, and Gemini 3.1—on a validated 80-item instrument (40 MCQs and 40 True/False statements) derived from 20 green building design scenarios spanning materials, energy, water, and cross-domain lifecycle trade-off reasoning. *Methods*: 80 items from 20 original scenarios were validated by two professors. MCQs required selection of the optimal design alternative under competing sustainability constraints; True/False items used context-dependent partial truths (Statements 1–20: False via contextual override; 21–40: True) to evaluate resistance to overgeneralization. Items were grounded in LEED v4.1, ASHRAE 90.1, and ISO 14044. Each model was evaluated under identical zero-temperature conditions using binary scoring. Pearson Chi-square tests were applied for inter-model and inter-format comparisons. *Results*: Claude achieved the highest overall accuracy (95.00%; 95% CI: 90.22–99.78%), followed by DeepSeek-R1 (93.75%), ChatGPT-5.2 (88.75%), and Gemini 3.1 (83.75%). No statistically significant difference across chatbots was found ($x^2 = 7.251$, df = 3, $P = 0.064$). No significant MCQ-vs-T/F differential was observed for any model (all $P > 0.05$); pooled T/F accuracy (91.3%) marginally exceeded pooled MCQ accuracy (89.4%). *Conclusions*: This is the first study to benchmark multiple frontier LLMs in green building design using a purpose-built, professor-validated instrument evaluating both design-alternative selection and resistance to absolute sustainability heuristics. Consistent performance declines in cross-domain lifecycle scenarios reveal residual multi-framework integration limitations across all models. Findings support LLM integration into green building education and decision-support, while underscoring the need for expert oversight in complex design contexts.

Keywords: Large language models; Green building design; Sustainable architecture; Benchmark evaluation; Energy efficiency; Water conservation; LCA; LEED; Claude; ChatGPT; DeepSeek; Gemini

Online publication: August 31, 2026

1. Introduction

Frontier large language models (LLMs), such as Claude, ChatGPT-5.2, DeepSeek-R1, and Gemini 3.1, can interpret complex technical prompts, reason across multi-criteria problems, and generate expert-like responses [1–4]. In architecture and engineering, LLMs address questions spanning energy modeling, building regulation, and sustainability with moderate-to-high accuracy, though performance varies by model and domain [5–7].

Green building design demands simultaneous integration of ASHRAE 90.1, water management, ISO 14044 Lifecycle Assessment (LCA), and Leadership in Energy and Environmental Design (LEED) v4.1 [8–12]. Systematic reviews catalog broad sustainability uses of LLMs without addressing validated, mixed-format instruments grounded in multi-domain lifecycle reasoning [13–14]. Practitioners must navigate scenarios where individually optimal domain choices conflict—e.g., thermally superior insulation with high embodied carbon, or on-site renewables vs. envelope improvement. A critical artificial intelligence (AI) failure mode is applying absolute heuristics (e.g., "local materials always have lower impact") without contextual adjustment. The True/False (T/F) component was engineered to target this failure: 20 False statements via contextual overrides and 20 True statements under established LCA principles.

No peer-reviewed study has evaluated LLMs on a professor-validated instrument combining MCQ design-alternative scenarios with high-difficulty True/False contextual partial-truth items across the full scope of green building domains (literature search: January 2023–August 2025) [5–7, 12]. This study aims to: (1) compare the overall accuracy of four frontier LLMs on 80 validated green building items; (2) assess performance differences by question format; and (3) identify domain-specific patterns of competency and limitation.

2. Methodology

2.1. Scenarios and questions

20 original green building scenarios were developed across four domains: materials and resources (Scenarios 1–5), energy efficiency (6–10), water conservation (11–15), and cross-domain lifecycle trade-off reasoning (16–20). Two MCQs per scenario (40 total; 4 options each) required integration of sustainability trade-offs rather than single-domain recall. 40 True/False items were constructed independently: Statements 1–20 were False via contextual override of general principles (all using "always"-assertions); Statements 21–40 were True under established LEED, ASHRAE, and LCA principles. Only text-based items were included.

2.2. Expert validation

Two professors independently assessed all 80 items against: (1) alignment with LEED v4.1, ASHRAE 90.1, and ISO 14044; (2) answer unambiguity within scenario context; (3) distractor plausibility / True/False contextual boundaries; and (4) difficulty appropriateness. Discordant ratings were resolved by structured consensus referencing primary technical standards.

2.3. Evaluation procedure

Each item was entered independently into all four chatbots using identical wording and format in fresh sessions to prevent prompt leakage. All models operated at temperature = 0. Three trials per item were conducted; modal responses were recorded for any inconsistency. Models were evaluated within a

consistent window to control for version confounds. The four LLMs represent diverse architectures: Claude (constitutional AI), ChatGPT-5.2 (RLHF), DeepSeek-R1 (mixture-of-experts with chain-of-thought pretraining), and Gemini 3.1 (Google DeepMind).

2.4. Assessment metrics

Binary scoring: correct = 1, incorrect = 0. Primary outcome: overall accuracy (80 items). Secondary: accuracy by format (MCQ vs. T/F). Pearson Chi-square tests for inter-model (x^2, df = 3) and intra-model format comparisons (x^2, df = 1). 95% CIs via the Wilson score method. Two-tailed; $P < 0.05$ threshold.

3. Results

Claude achieved the highest overall accuracy (76/80; 95.00%; 95% CI: 90.22–99.78%). DeepSeek-R1 followed (75/80; 93.75%), then ChatGPT-5.2 (71/80; 88.75%), and Gemini 3.1 (67/80; 83.75%), as shown in **Table 1**. Overall performance spread: 11.25 percentage points.

Table 1. Accuracy of AI chatbots (n = 80 items per model)

Model	Correct (of 80)	Accuracy (%)	SD	95% CI (%)
Claude	76	95.00	0.218	90.22–99.78
DeepSeek-R1	75	93.75	0.242	88.44–99.06
ChatGPT-5.2	71	88.75	0.316	81.83–95.67
Gemini 3.1	67	83.75	0.369	75.66–91.84

Note: SD = standard deviation. 95% CI = Wilson score confidence interval.

Pearson Chi-square across all chatbots: $x^2(3) = 7.251$, $P = 0.064$. The marginal non-significance is likely attributable to a moderate sample size (n = 80/model). All four models substantially exceeded random baselines (25% MCQ, 50% T/F).

Format analysis: Claude achieved 92.5% MCQs / 97.5% T/F ($P = 0.305$); DeepSeek-R1 92.5% / 95.0% ($P = 0.644$); ChatGPT-5.2 87.5% / 90.0% ($P = 0.723$); Gemini 3.1 85.0% / 82.5% ($P = 0.762$), as shown in **Table 2**. No significant MCQ-vs-T/F differential for any model (all $P > 0.05$).

Table 2. Performance by question format per chatbot

Model	MCQ Correct	MCQ (%)	T/F Correct	T/F (%)	Diff. (pp)	*P*-value
Claude	37/40	92.5	39/40	97.5	+5.0	0.305
DeepSeek-R1	37/40	92.5	38/40	95.0	+2.5	0.644
ChatGPT-5.2	35/40	87.5	36/40	90.0	+2.5	0.723
Gemini 3.1	34/40	85.0	33/40	82.5	-2.5	0.762

Note: T/F = True/False. All $P > 0.05$.

Pooled across all models: 143/160 MCQ responses correct (89.4%) and 146/160 T/F correct (91.3%). Three of four models performed marginally better on T/F than MCQ—indicating the high-difficulty True/False format did not systematically disadvantage models. Gemini 3.1 was the sole exception (T/F 82.5% vs. MCQ 85.0%), potentially indicating greater sensitivity to the contextual partial-truth framing.

All models performed strongest in materials and resources (Scenarios 1–5), with consistent accuracy decline through energy, water, and cross-domain scenarios. Cross-domain lifecycle trade-off scenarios (16–20) were the most challenging; Gemini 3.1 dropped to ~70%. Performance ranking—Claude > DeepSeek-R1 > ChatGPT-5.2 > Gemini 3.1—was stable across all domains.

4. Discussion

All four chatbots performed substantially above chance, with Claude achieving 95.00% and all models exceeding 83%, indicating that frontier LLMs encode meaningful LEED-aligned green building knowledge and can apply it reliably in structured text-based scenarios [15–18]. Performance declined progressively from materials through energy, water, to cross-domain scenarios—consistent with increasing need for simultaneous integration of multiple technical frameworks—and this domain-ordering pattern was stable across all models.

The absence of significant MCQ-vs-T/F differentials for any model validates the mixed-format design, confirming both item types assess a consistent underlying competency construct. That three of four models performed marginally better on T/F (despite elevated design difficulty) indicates the contextual partial-truth format successfully assessed domain knowledge rather than a separate format-dependent skill. Gemini 3.1's slight T/F underperformance may reflect differential sensitivity to "always"-assertion constructions.

The professor validation process—two independent LEED AP BD+C-accredited experts reviewing all 80 items against LEED v4.1, ASHRAE 90.1, and ISO 14044, with structured consensus for discordant ratings—substantially strengthens internal validity over prior LLM sustainability evaluations using unvalidated question sets [5, 7]. DeepSeek-R1's high performance (93.75%) may reflect its mixture-of-experts (MoE) architecture and chain-of-thought pretraining supporting multi-framework integration. Gemini 3.1's wider CI (75.66–91.84%) suggests less stable green building knowledge encoding.

Limitations: no human comparison group (LEED practitioners or students); exclusion of graphical content; risk that performance partly reflects training-data recognition. Post-hoc power analysis estimates 120–150 items per model needed for 80% power to detect the observed 11.25 pp differential at $\alpha = 0.05$. Future work should incorporate human benchmarks, diverse validated sources, and retrieval-augmented generation (RAG) coupling LLMs with structured green building knowledge bases.

5. Conclusion

Claude, DeepSeek-R1, ChatGPT-5.2, and Gemini 3.1 can accurately answer professor-validated, case-based green building design items across energy efficiency, water conservation, materials, and cross-domain lifecycle trade-off reasoning, with overall accuracies of 83.75%–95.00% across an 80-item instrument. This is the first study to benchmark multiple frontier LLMs in green building design using a purpose-built, professor-validated instrument simultaneously evaluating design-alternative selection and resistance to absolute sustainability heuristics. Although no statistically significant overall difference was found ($x^2 = 7.251$, df = 3, $P = 0.064$), and no significant MCQ-vs-T/F differential was detected for any model (all $P > 0.05$), consistent performance declines in cross-domain lifecycle scenarios reveal residual multi-framework integration limitations. These findings support further validation, larger datasets, and expert human oversight before autonomous LLM deployment in high-stakes sustainable building design contexts.

Disclosure statement

The authors declare no conflict of interest.

References

[1] Biswas SS, 2023, Role of ChatGPT in public health. Annals of Biomedical Engineering, 2023(51): 868–869.

[2] Ali SR, Dobbs TD, Hutchings HA, et al., 2023, Using ChatGPT to Write Patient Clinic Letters. The Lancet Digital Health, 2023(5): e179–181.

[3] Wang Y, Liang L, Li R, et al., 2024, Comparison of ChatGPT, Claude, and Bard in Support of Technical Decision-making. Journal of Multidisciplinary Healthcare, 2024(17): 3917–3929.

[4] DeepSeek-AI, 2025, DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv: 2501.12948.

[5] Wu J, Jiang M, Fan J, et al., 2025, Arch-Eval Benchmark for Assessing Chinese Architectural Domain Knowledge in LLMs. Scientific Reports, 2025(15): 1–12.

[6] He Q, 2024, How Well Do LLMs Perform in Green Building Assessment? iCCPMCE-2024.

[7] Nguyen HC, Dang HP, Nguyen TL, et al., 2025, Accuracy of Latest LLMs in Answering MCQs in Technical Domains. PLoS ONE, 2025(20): e0317423.

[8] U.S. Green Building Council, 2021, LEED v4.1 Building Design and Construction Reference Guide. USGBC; 2021.

[9] Lu C, Li S, Lu Z, 2021, Building Energy Prediction Using Artificial Neural Networks: A Literature Survey. Energy and Buildings, 2021(233): 110919.

[10] Luo X, Zhang Y, Lu J, et al., 2024, Multi-objective Optimization of Office Park Building Envelope for Nearly Zero Energy. Journal of Building Engineering, 2024(86): 108713.

[11] Yu L, Sun Y, Xu Z, et al., 2021, Multi-agent Deep Reinforcement Learning for HVAC Control. IEEE Transactions on Smart Grid, 12(1): 407–419.

[12] Zhong S, Aseniero BA, Groom AI, et al., 2025, Towards Interactive AI-assisted Material selection for Sustainable Building Design. Companion Publication of the 2025 ACM Designing Interactive Systems Conference, 567–573.

[13] Raeissi MM, Knapen R, 2025, Applications of Generative LLMs in Environmental Science: A Systematic Review. Advances in Environmental and Engineering Research, 6(2): 1–20.

[14] Gao Y, Yiu TW, Shen X, et al., 2026, LLMs in Smart Construction: A Systematic Review. Engineering, Construction and Architectural Management, 33(15): 159–181. https://doi.org/10.1108/ECAM-12-2024-1402

[15] Center for AI Safety, Scale AI, HLE Contributors Consortium, 2026, A Benchmark of Expert-level Academic Questions to Assess AI Capabilities. Nature, 2026(649): 1139–1146.

[16] Shen Y, Heacock L, Elias J, et al., 2023, ChatGPT and other LLMs are Double-Edged Swords. Radiology, 307(2): e230163. https://doi.org/10.1148/RADIOL.230163

[17] Kambhampati S, 2024, Can LLMs Plan? The Science and Nonsense of LLM Reasoning Claims. Communications of the ACM, 67(4): 34–41.

[18] Turpin M, Michael J, Perez E, et al., 2023, Language Models Don't Always Say What They Think. Advances in Neural Information Processing Systems, 2023(36): 74952–74965.