Developments in the Built Environment 27 (2026) 101007

Contents lists available at ScienceDirect

Developments in the Built Environment

journal homepage: www.sciencedirect.com/journal/developments-in-the-built-environment

Ethical assessment of large language model-generated advisory text on designing the built environment for health

Mohammad Javad Koohsari [a,b,c], Becky P.Y. Loo [d], Jing Zhao [e,*], Jiuling Li [a], Ying Long [f], Yi Lu [g], Koichiro Oka [b], Andrew T. Kaczynski [c]

[a] Urban Design Science for Health Laboratory, Japan Advanced Institute of Science and Technology, Japan
[b] Faculty of Sport Sciences, Waseda University, Japan
[c] Department of Health Promotion Education and Behavior, Arnold School of Public Health, University of South Carolina, Columbia, United States
[d] Department of Geography, The University of Hong Kong, Hong Kong, China
[e] School of Architecture and Urban Planning, Guangzhou University, China
[f] School of Architecture, Tsinghua University, China
[g] Department of Architecture and Civil Engineering, City University of Hong Kong, Hong Kong, China

ARTICLE INFO

Keywords:
Urban design ethics
Artificial intelligence (AI)
Urban design science
Healthy design
Urban form

ABSTRACT

Large language models (LLMs) can generate advisory text on modifying built environments to support health. This study examined the ethical properties of a recent LLM generating text on built environments to support health. The prompts covered six health-related pathways in higher-income, lower-income, and mixed-income neighbourhoods. Overall, 180 answers were coded against four ethical criteria. Non-maleficence was satisfied in all answers. Lower-income contexts were rarely offered weaker proposals than higher-income contexts. Reference to collective participation and transparent oversight appeared in 70-90% of answers without a budget constraint, but only 30-50% under one. The LLM more consistently met minimum expectations for harm avoidance and distributive justice than for collective participation and transparent oversight. These findings suggest that current LLM outputs may reproduce some baseline ethical conventions in urban design discourse but are less reliable on procedural concerns. LLM-generated outputs should therefore be interpreted cautiously within existing built environment decision-making processes.

1. Introduction

Interest in urban design as a means to improve population health has revived over the last decade (Koohsari and Kaczynski, 2025). This renewed focus reflects growing concerns about non-communicable diseases, social isolation, and health inequities in urban settings (Cacciatore et al., 2025). Urban design shapes the built environment that residents experience every day through decisions about density, land use, and public spaces, among others. The built environment is a modifiable factor in health that operates through behavioural (e.g., physical activity, diet) and exposure (e.g., air pollution, heat) pathways (Frank et al., 2019; Rajagopalan et al., 2024). Evidence on these mechanisms has expanded across epidemiology, public health, and urban studies. In parallel, research in this field has increasingly used digital tools, such as geographic information systems and simulation models, to examine relationships between the built environment and health outcomes (Dheekwal et al., 2025; Stankov et al., 2020).

Large language models (LLMs) are a class of artificial intelligence (AI) systems that generate text in response to open-ended prompts (Brown et al., 2020). LLMs are already used in several fields related to health and place, including urban studies, healthcare, and clinical medicine (Bedi et al., 2025; Thirunavukarasu et al., 2023; Xia et al., 2025). In the urban design and planning field, recent studies have used LLMs to support participatory planning (Zhang et al., 2025a), evaluate urban plans (Fu et al., 2024), examine zoning codes (Salazar-Miranda and Talen, 2025), and assist with sustainability compliance for high-rise construction (Wong and Loo, 2025). In clinical and public health contexts, LLMs have been explored for drafting patient summaries (Van Veen et al., 2024), supporting diagnostic reasoning (Goh et al., 2024), and detecting disease outbreaks (Kwok et al., 2024).

* Corresponding author. 230 Wai Huan Xi Road, Guangzhou Higher Education Mega Center, Guangzhou, Guangdong Province, 510006, China.
E-mail addresses: koohsari@jaist.ac.jp (M.J. Koohsari), bpyloo@hku.hk (B.P.Y. Loo), zhaojing@gzhu.edu.cn (J. Zhao), li-jiuling@jaist.ac.jp (J. Li), ylong@tsinghua.edu.cn (Y. Long), yilu24@cityu.edu.hk (Y. Lu), koka@waseda.jp (K. Oka), atkaczyn@mailbox.sc.edu (A.T. Kaczynski).

https://doi.org/10.1016/j.dibe.2026.101007
Received 7 January 2026; Received in revised form 12 April 2026; Accepted 5 August 2026
Available online 10 August 2026

Because LLMs can respond to domain-specific prompts in natural language, they can generate prescriptive text that states what should be done in each situation rather than only providing descriptive summaries (Thirunavukarasu et al., 2023; Wu et al., 2024). Their outputs can imitate the language and structure of expert advice and may give the impression of expert judgement (Bender et al., 2021; Messeri and Crockett, 2024). For example, several recent studies have shown that LLMs can turn short patient vignettes into concise proposals for diagnostic steps (Goh et al., 2024; Qiu et al., 2025). Similarly, LLMs can be prompted to provide advisory text on how the built environment could be modified to improve health (Koohsari et al., 2025). This has practical relevance for the built environment field, where such outputs may be assumed to reflect expert judgement.

Recent work in urban planning and clinical medicine has raised ethical concerns about the use of LLMs in decision-making contexts (Haltaufderheide and Ranisch, 2024; Sanchez et al., 2025). For instance, using AI tools to inform urban planning decisions can evoke ethical issues such as bias, socio-spatial inequity, privacy risk, and reduced transparency (Sanchez et al., 2025). In clinical medicine, some studies on LLM ethics have identified concerns about fairness, non-maleficence, transparency, privacy, and the production of convincing but inaccurate information (Haltaufderheide and Ranisch, 2024). The extant literature suggests that LLMs should not be treated as neutral tools and that their outputs constitute ethically laden recommendations with potentially unintended consequences for health (Ong et al., 2024b). Some empirical studies in clinical medicine have begun to assess these ethical issues when evaluating LLM advice (Esmaeilzadeh, 2025; Zack et al., 2024). For instance, Zack et al. (2024) evaluated GPT-4, an LLM related to ChatGPT, using clinical vignettes to test whether diagnoses and management plans differed by patient race and gender. They found systematic race- and gender-based differences in both diagnoses and management (Zack et al., 2024). These studies show that LLM outputs can be examined empirically for ethical variation across contexts. They also highlight that the relevant ethical concerns differ across fields of use. In built environment research, however, ethical judgement is rarely framed only in terms of harm avoidance or fairness of outcomes. It is intertwined with public interest, procedural justice in planning, the politics of participation, and the legitimacy of expert authority in shaping urban futures (Matamanda et al., 2025; McKay et al., 2011). In this context, LLM-generated advice enters a field in which ethical judgement is historically negotiated through planning institutions, professional norms, and design practice rather than being assessed only as a technical output. However, ethical questions about LLM-generated guidance for the built environment and health have received little in-depth investigation. This is also a gap for built environment practice because LLMs are increasingly used in planning-related contexts without clear evidence on how their outputs address ethical concerns.

To address these gaps, the current study empirically examines the ethical properties of advice generated by an LLM on how the built environment can be modified to support human health. The analysis draws on ethical concerns identified in domains of urban design and clinical medicine. The study design was guided by prior literature on built environment and health, and by recent works on LLM ethics and urban planning. Unlike prior empirical studies that have examined ethical issues in clinical contexts, this study focuses on LLM-generated built environment advice for health. The study asked whether LLM-generated advice on modifying the built environment for health met minimum ethical expectations across different neighbourhood income contexts and budget conditions.

2. Methods

2.1. Study design

This study used a quasi-experimental design to evaluate the behaviour of an LLM when prompted to generate advisory text on urban design and health. Prompts were constructed to obtain advice from an LLM about how the built environment could be changed to support human health. The LLM was implemented using ChatGPT, a widely used conversational system. ChatGPT is a general-purpose LLM developed by OpenAI and made available to the public in 2022. In this study, ChatGPT was treated as a source of quasi-expert advisory text on questions about the built environment and health. The study did not build, train, or fine-tune a model. Instead, it evaluated repeated outputs from an existing model under predefined prompt conditions. As elaborated below, experimental conditions were defined by variations in health pathways and neighbourhood income context, and each distinct combination produced one or more text outputs.

2.2. Selection of health pathways and neighbourhood contexts

Health pathways were based on behavioural and exposure mechanisms linking the built environment with health proposed by Frank et al. (2019). Behavioural pathways included physical activity, dietary intake, and social interaction. Exposure pathways covered air pollution, traffic safety and crime, and noise. These pathways were selected to capture established mechanisms linking built environment features with health. They also provided a theory-informed basis for examining LLM-generated outputs under different health-related prompt conditions. Two relatively homogenous urban neighbourhoods were first specified, one higher-income and one lower-income. A third context described a mixed-income urban neighbourhood with both higher-income and lower-income residential blocks. These income contrasts were included to allow comparisons of the ethical properties of LLM advice across different socioeconomic conditions.

2.3. Prompt development

Prompt texts were designed to obtain comparable advice on the built environment and health across all conditions. All prompts began with the same role instruction: “You are an urban design and population health adviser.” Two sets of prompts were created. The first set comprised 12 prompts, one for each combination of the six health pathways and the two neighbourhood income contexts. Each prompt included a brief description of the neighbourhood with an income descriptor (“a higher-income neighbourhood of a large city” or “a lower-income neighbourhood of a large city”) and a short problem statement aligned with the relevant behavioural or exposure pathway. The main instruction asked ChatGPT to propose changes to the built environment and urban design regulations. Guidance directed responses towards four aspects of the built environment, identified by Frank et al. (2019): transportation infrastructure, land use/walkability, pedestrian environment, and greenspace. Prompts explicitly asked not to suggest social programmes, health education campaigns, clinical care, or general taxation policies. Each prompt included a brief cue encouraging human oversight. It also asked the model to note when advice might be unsafe or too context dependent and to defer final decisions to local professionals and residents. Answers were restricted to 60-80 words and were requested to focus on the most important physical changes. The full wording of all prompts is provided in Supplementary A.

The second set included six prompts, one per health pathway, and described a single mixed-income neighbourhood with both higher- and lower-income residential blocks (Supplementary A). The same role instruction, problem framing, scope guidance, exclusions, humility cue, and word limit were used. An additional element introduced was a budget constraint, which stated that city funding cannot support major changes in all parts of the neighbourhood at once. Prompts asked ChatGPT to indicate which blocks or streets should be prioritised for the biggest change and to explain the reasoning behind the prioritisation.

2.4. LLM configuration and data collection

ChatGPT was accessed through the standard web interface. The active model version at the time of data collection was GPT-5.1. Data were collected on a single day in November 2025 to limit temporal variations in model behaviour. Chat history was turned off before data collection. Each prompt was entered in a new chat with no previous messages. No follow-up questions, regenerations, or additional system instructions were used. Default interface settings were applied, with no manual changes made to sampling parameters. Each of the 18 prompts was run 10 times to capture variations in outputs. This follows recent studies that sample multiple generations per input when evaluating LLM behaviour (Gallo et al., 2025; Gu et al., 2025). In the present study, repeated runs were used to characterise variation in generated outputs under the same prompt conditions. The 12 prompts in set 1 and the six prompts in set 2 generated 120 and 60 responses, respectively.

2.5. Ethical evaluation criteria

Each advice text was evaluated on four ethical criteria: non-maleficence, distributive justice across socio-spatial contexts (hereafter "distributive justice"), collective autonomy and participation (hereafter "collective participation"), and transparency, epistemic humility, and human oversight (hereafter "transparent oversight"). These criteria were selected based on concerns in LLM ethics in clinical medicine and in AI applications in urban planning about safety, distributive fairness, public involvement, and transparency (Haltaufderheide and Ranisch, 2024; Sanchez et al., 2025). They were also chosen to reflect established concerns in built environment scholarship about how spatial decisions are justified, whose interests are represented, and how expert authority is constrained through public and professional scrutiny (Leyden et al., 2017).

Non-maleficence assessed whether advice avoided clearly harmful or unsafe built environment changes that would worsen the health exposure or risk named in the question. In built environment terms, this criterion was treated as a minimum safety screen and reflects a basic professional baseline rather than a complete ethical judgement. Distributive justice across socio-spatial contexts evaluated whether lower-income neighbourhoods or blocks were treated at least as favourably as higher-income neighbourhoods or blocks. This operationalisation is closer to formal equality than to the equity-based reasoning that dominates built environment debates (Seyedrezaei et al., 2023). Collective participation considered whether residents or communities were recognised as decision-making agents, for example through consultation or co-design. Transparent oversight captured whether answers acknowledged uncertainty or context dependence or indicated that decisions should be reviewed by local professionals or residents. In this analysis, non-maleficence and distributive justice served as outcome-oriented ethical criteria, and collective participation and transparent oversight were procedural ethical criteria (Devon and Van de Poel, 2004). Each criterion was coded as a binary outcome (1 = criterion satisfied, 0 = criterion not satisfied). All 180 answers were coded independently by two co-authors (JL and JZ), both with training in urban design and public health. Coding was based on the written answers using a predefined coding protocol and example (Supplementary B). Disagreements were resolved by discussion until consensus was reached.

2.6. Statistical analysis

Descriptive statistics were used to summarise the four ethical criteria for both prompt sets. For set 1, the unit of analysis was the individual answer, defined as a single LLM output to a single prompt. Proportions of answers satisfying each criterion were calculated for all 120 answers and stratified by neighbourhood income (higher-income versus lower-income). To examine distributive justice in set 1, the unit of analysis was the paired higher-income and lower-income answers for the same pathway and run, giving 60 answer pairs in total (6 pathways × 10 runs). The proportion of pairs satisfying distributive justice was estimated with 95% confidence intervals (95% CIs), and an exact binomial test with a null value of 0.50 was used as an exploratory check to assess whether the observed proportion differed from chance. Within-pair differences in collective participation and transparent oversight were described by classifying each pair into four categories: both answers satisfied the criterion, neither satisfied it, only the higher-income answer satisfied it, or only the lower-income answer satisfied it. McNemar's test was applied to discordant pairs.

For set 2, the unit of analysis was the individual answer (60 answers in total). Proportions and 95% CIs were calculated for all four criteria. For the allocation-based distributive justice outcome, an exact binomial test with a null value of 0.50 was applied. To examine differences in ethical performance between the two prompt sets, secondary analyses compared non-maleficence, collective participation, and transparent oversight between sets 1 and 2 using Pearson χ^2 tests on two-by-two tables. Because each prompt was run 10 times, answers from the same prompt represent repeated observations rather than independent data points. Proportions and CIs are therefore presented as descriptive summaries of model behaviour under the specified prompts, and all hypothesis tests were interpreted as exploratory.

To assess the influence of repeated runs per prompt, a sensitivity analysis was undertaken using prompts as the unit of analysis. For each prompt, a majority-rule indicator was created for non-maleficence, collective participation, and transparent oversight. A majority-rule indicator was also created for distributive justice, using higher-lower income neighbourhood pairs in set 1 and allocation decisions in set 2. These indicators were coded as satisfied if at least eight of ten runs met the criterion. Key descriptive summaries and income contrasts were recomputed using these prompt-based indicators to assess whether repeated runs per prompt altered the overall pattern of results. All analyses were conducted using SPSS 29 (SPSS Inc., Chicago, IL).

3. Results

3.1. Overview of advice texts and coding outcomes

The two prompt sets generated 180 advice texts in total (120 in set 1 and 60 in set 2). Non-maleficence was satisfied in 180 of 180 answers (100.0%). Distributive justice was evaluable for 60 paired answers in set 1 and for 60 answers in set 2; overall, 110 of 120 evaluable units (91.7%) satisfied this criterion. Collective participation was coded in 109 of 180 answers (60.6%). Transparent oversight cues were present in 136 of 180 answers (75.6%). Frequencies for non-maleficence, collective participation, and transparent oversight by prompt set and neighbourhood income context are summarised in Table 1. The distributive justice criterion for set 1 was assessed using higher-lower income answer pairs and is presented separately in Table 2.

Table 1
Ethical coding outcomes for individual answers by neighbourhood income context and prompt set.

Prompt set and context	n	Non-maleficence n (%)	Collective participation n (%)	Transparent oversight n (%)
All answers	180	180 (100)	109 (60.6)	136 (75.6)
Set 1 - higher-income neighbourhood	60	60 (100.0)	40 (66.7)	50 (83.3)
Set 1 - lower-income neighbourhood	60	60 (100.0)	48 (80.0)	58 (96.7)
Set 2 - mixed-income, budget-constrained neighbourhood	60	60 (100.0)	21 (35.0)	28 (46.7)

Table 2
Distributive justice outcomes by prompt set.

Prompt set	n	Distributive justice satisfied n (%)	95% CI	Exact binomial p-value n (%)
Set 1 - paired answers	60	51 (85.0)	[76.0, 94.0]	$p < 0.001$
Set 2 - mixed-income, budget-constrained neighbourhood	60	59 (98.3)	[95.0, 100.0]	$p < 0.001$

3.2. Ethical properties of advice across neighbourhood income contexts

Set 1 included 60 answers for higher-income and 60 answers for lower-income neighbourhoods. As shown in Table 1, non-maleficence was satisfied in 60 of 60 higher-income answers (100.0%) and 60 of 60 lower-income answers (100.0%). Collective participation was coded in 40 of 60 higher-income answers (66.7%; 95% CI [54.0, 79.0]) and 48 of 60 lower-income answers (80.0%; 95% CI [70.0, 90.0]). Transparent oversight was coded in 50 of 60 higher-income answers (83.3%; 95% CI [74.0, 93.0]) and 58 of 60 lower-income answers (96.7%; 95% CI [92.0, 100.0]).

Paired analyses used the 60 higher–lower income neighbourhood pairs. Non-maleficence was satisfied in both answers in all 60 pairs (100.0%). For distributive justice, the criterion was satisfied in 51 of 60 pairs (85.0%; 95% CI [76.0, 94.0]; exact binomial test $p < 0.001$). For collective participation, 32 pairs (53.3%) showed participation in both answers, four pairs (6.7%) in neither answer, eight pairs (13.3%) only in the higher-income answer, and 16 pairs (26.7%) only in the lower-income answer (McNemar's test $p = 0.15$). For transparent oversight, 49 pairs (81.7%) showed transparent oversight cues in both answers, one pair (1.7%) in neither answer, one pair (1.7%) only in the higher-income answer, and nine pairs (15.0%) only in the lower-income answer (McNemar's test $p = 0.02$) (Tables 2 and 3).

3.3. Ethical properties of advice under budget-constrained prioritisation

Set 2 comprised 60 answers for a mixed-income neighbourhood in which major built environment changes had to be prioritised under a budget constraint. As shown in Table 1, non-maleficence was satisfied in 60 of 60 answers (100.0%). Distributive justice was satisfied in 59 of 60 answers (98.3%; 95% CI [95.0, 100.0]; exact binomial test $p < 0.001$) (Table 2). Collective participation was coded in 21 of 60 answers (35.0%; 95% CI [23.0, 47.0]) (Table 1). Transparent oversight was coded in 28 of 60 answers (46.7%; 95% CI [34.0, 60.0]) (Table 1).

3.4. Comparison between prompt sets

We also compared responses between prompt sets to examine potential effects of including a budget constraint in the prompt (set 2). Across all 180 answers, non-maleficence was satisfied in 100.0% of set 1 answers and 100.0% of set 2 answers. Collective participation was coded in 73.3% of answers in set 1 and 35.0% in set 2 ($\chi^2 = 24.6$, $p < 0.001$). Transparent oversight was coded in 90.0% of answers in set 1 and 46.7% in set 2 ($\chi^2 = 40.7$, $p < 0.001$). These between-set comparisons were treated as exploratory because multiple answers were generated for each underlying prompt. Sensitivity analyses that aggregated the 10 runs for each prompt into majority-rule indicators produced proportions and contrasts consistent with those based on individual answers. This suggests that clustering of runs within the prompt did not affect the main pattern of findings.

4. Discussion

4.1. Main findings

This study examined the ethical properties of advice generated by an LLM, ChatGPT, when it was asked to propose modifications to the built environment to support residents' health. The prompts covered different neighbourhood income contexts and budget conditions. The assessment drew on four shared ethical concerns in urban planning and medicine, including non-maleficence, distributive justice, collective participation, and transparent oversight (Haltaufderheide and Ranisch, 2024; Sanchez et al., 2025). Across these criteria, ChatGPT showed a mixed pattern. It consistently avoided clearly harmful recommendations and did not disadvantage lower-income neighbourhoods. In contrast, it paid less attention to procedural ethical concerns such as involving residents and acknowledging uncertainty. These findings are important in built environment research where ethical judgement is shaped by both the content of spatial proposals and the ways decisions are justified, scrutinised, and made legitimate through planning processes (Durrant and Kossberg, 2023; Watson, 2006). Similar procedural ethical concerns, including limited stakeholder participation, transparency of limitations, and human oversight, have been identified in the application of LLMs in urban planning (Raymond et al., 2025; Zheng et al., 2025). A recent overview of ethical challenges of LLMs in medicine also highlighted substantial concerns about limited transparency of model reasoning, the risk of hallucinated content, and the need for human oversight (Ong et al., 2024a). To our knowledge, this is the first ethical assessment of LLM-generated urban design advice for promoting residents' health.

4.2. Non-maleficence and distributive justice

This study found that the non-maleficence criterion was satisfied in all answers, across neighbourhood income contexts and budget conditions. Recent LLMs, including ChatGPT, undergo safety alignment using reinforcement learning from human feedback, which discourages obviously harmful content (Ouyang et al., 2022). The urban design and public health literature has long treated the built environment as a means of reducing safety risks, such as traffic crashes and air pollution (Dannenberg et al., 2003; Frumkin, 2002). These norms are likely reflected in LLMs' training data. In this study, the prompts assigned the LLM the role of an urban design and population health adviser and asked it to advise on how to improve residents' health. This framing is close to professional expectations in urban design and public health and likely encouraged advisory responses that avoided obvious harms. In the context of built environment research, however, this finding should be interpreted as meeting a minimum professional safety baseline rather than as a complete ethical judgement. At the same time, non-maleficence was coded as a binary judgement and could not detect

Table 3
Collective participation and transparent oversight patterns for higher-lower income neighbourhood pairs in set 1 (n = 60).

Criterion	Both answers n (%)	Neither answer n (%)	Higher-income only n (%)	Lower-income only n (%)	McNemar p-value
Collective participation	32 (53.3)	4 (6.7)	8 (13.3)	16 (26.7)	0.15
Transparent oversight	49 (81.7)	1 (1.7)	1 (1.7)	9 (15.0)	0.02

Note. Each row summarises 60 pairs of answers, one for a higher-income neighbourhood and one for a lower-income neighbourhood. "Both answers" means the criterion was present in both answers in the pair. "Neither answer" means it was absent in both. "Higher-income only" and "lower-income only" mean that the criterion was present only in the answer for the higher-income or lower-income neighbourhood, respectively. McNemar p-values test the difference between the "higher-income only" and "lower-income only" categories.

subtle or long-term risks. These results suggest that health-oriented urban design prompts can direct LLMs to function as a basic content filter against clearly hazardous built environment proposals. However, this is not a sufficient condition for ethically robust or health-promoting design actions, since the non-maleficence criterion, as coded in this study, does not consider context-specific situations, longer-term risks, or wider political and distributive consequences. For example, improving walkability in a lower-income neighbourhood may appear to be safe advice. Nevertheless, such changes can contribute to housing unaffordability (Christie et al., 2023; Koschinsky and Talen, 2015), and, over time, to the displacement of poorer residents to areas with even less health-supportive environments. In practice, all urban design advice generated by LLMs still requires safety assessment and broader ethical review by trained urban designers and public health professionals. Future studies are needed to test LLMs with more demanding prompts to assess place-based safety over longer periods.

In this study, the distributive justice criterion also showed strong performance across both prompt sets. In each higher-lower income neighbourhood pair in set 1, and for the lower-income blocks within the mixed-income neighbourhood in set 2, lower-income neighbourhood contexts were rarely treated less favourably than their higher-income counterparts in the extracted advice. Clear labelling of neighbourhood income contexts in the prompts may have supported this pattern. Current evidence in the built environment and health seldom specifies or evaluates distinct "gold standard" physical designs for higher- and lower-income neighbourhoods that face the same public health issue (Smith et al., 2017). An LLM trained on this literature is therefore unlikely to invent income-specific built environment changes to promote health for different income contexts. Equity-oriented urban and transportation planning scholarship, however, often argues for prioritised resource allocation to disadvantaged areas to address their historic underinvestment (Krapp et al., 2021; Loh and Kim, 2021). The present study adopted a narrower ethical expectation of distributive justice focused on a minimum standard: it checked that lower-income neighbourhoods did not receive weaker built environment proposals than higher-income neighbourhoods facing the same health problem. Strong performance on this criterion indicates that, under the prompts used here, the model did not produce a simple income-based disadvantage in its written advice. It does not, however, show that the model reasoned in ways that would prioritise disadvantaged areas in response to greater need or historic underinvestment. For urban design and health practice, this suggests that, with similar prompts, LLM-generated advice may help avoid obvious downgrading of solutions for disadvantaged areas. However, the present study cannot test whether the scale and pace of recommended built environment changes are sufficient to reduce the observed health and resource gaps between different areas. Future work in urban design and public health research should therefore test prompts and equity indicators that reflect baseline conditions and constraints of existing neighbourhoods.

4.3. Practical and social implications

In this study, collective participation and transparent oversight criteria were present in many answers, but less frequently than non-maleficence and distributive justice. Meeting the procedural ethics criteria in this study required clear references to residents, stakeholders, uncertainty, or human oversight. Studies show that current LLMs tend to produce confident technical outputs and fail to appropriately express uncertainty that might necessitate participation and oversight (Griot et al., 2025; Steyvers et al., 2025). The present study also found that the proportion of answers satisfying these procedural criteria declined sharply in the budget-constrained prioritisation prompts (set 2) compared with the prompts that did not include a budget constraint (set 1). In the budget-constrained prompts, the prompts framed the task as selecting and justifying a limited set of built environment changes under a shared budget constraint (e.g., which street design improvements should be prioritised). In this setting, ChatGPT tended to use the available answer space to specify which built environment actions should be funded and why, leaving less room to mention residents or oversight. This is consistent with evidence that LLM behaviour is highly sensitive to prompt framing (Germani and Spitale, 2025; Zhang et al., 2025b). These findings have direct implications for the built environment and health field. Urban planning research has long recognised tensions between participation and decision efficiency, especially when limited resources require decisions about priority (Calderon et al., 2024; Kahila and Kyttä, 2009). The decline in participation and oversight under budget constraints may suggest a more technocratic pattern in the model's outputs. Under conditions of scarcity, built environment decisions still involve questions of public interest, legitimacy, and the role of expertise (Araos, 2023; Herkes and Zouridis, 2023). The relative absence of community participation in LLM-generated advice for urban design for health may lead to built environment recommendations without contextual understanding or sufficient democratic legitimacy. Future studies should test prompting strategies and protocols that foreground residents' feedback when LLMs support urban design decisions for health. Relatedly, ChatGPT was also inconsistent about flagging "what it did not know". This may lead to decision-makers' overconfidence about the validity of LLM-based advice in designing neighbourhoods for health. Comparable concerns have been reported in the construction sector by Wong and Loo (2025). They found that domain-specific LLM systems for sustainability compliance in high-rise construction still required a "human--in-the-loop" review protocol to prevent misinterpretation of regulatory standards. Similar to work in other fields (Albadarin et al., 2024; Lucas et al., 2024), developing frameworks to include a "human-in-the-loop" for LLM-based built environment and health decisions remains necessary. Therefore, LLM-generated outputs are best understood as advisory inputs within urban design and planning processes, which require professional judgement, participatory scrutiny, and institutional review.

4.4. Limitations and future research

This study has several limitations. Only one LLM (ChatGPT 5.1) was assessed through the standard web interface on a single day with default settings. This limits generalisability to other LLMs and cannot account for version updates or configuration changes. Ethical properties were coded as binary outcomes, which may not capture finer distinctions. Advice texts were also limited to 60-80 words, which may have constrained participation and oversight language. Additionally, this study evaluated written recommendations under controlled prompt conditions rather than real-world planning processes. The findings therefore should not be interpreted as evidence of how LLMs would perform within actual planning institutions, consultation processes, or statutory decision-making. Given that digital technologies are expected to play an increasing role in future built environment and health research and practice, future studies should test these ethical dimensions when LLMs are used in real-world decision-making settings, such as neighbourhood redevelopment and urban regeneration, to support health-supportive built environments.

5. Conclusions

In this study, LLM-generated advice on urban design for health met some, but not all, of the ethical expectations examined under the specific prompts and tasks used. The advice did not clearly propose harmful built environment changes and did not show a simple income-based disadvantage in its recommendations. In contrast, references to collective participation and transparent oversight were less frequent and declined further in the budget-constrained prompts, which are most relevant to real-world urban design decision-making about neighbourhood prioritisation. These findings suggest that current LLM outputs may reflect some baseline ethical expectations in built environment practice but are less consistent in addressing the procedural concerns that are central to

urban planning ethics. Therefore, their use in urban design for health requires careful interpretation within existing professional, participatory, and institutional processes. One remaining question for future research is how changes in prompts, model design, and governance arrangements influence the ethical performance of LLM-generated advice on designing built environments to support health.

CRediT authorship contribution statement

Mohammad Javad Koohsari: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Supervision, Validation, Writing – original draft, Writing – review & editing. **Becky P.Y. Loo:** Investigation, Methodology, Validation, Writing – review & editing. **Jing Zhao:** Investigation, Methodology, Writing – review & editing. **Jiuling Li:** Investigation, Methodology, Writing – review & editing. **Ying Long:** Writing – review & editing. **Yi Lu:** Writing – review & editing. **Koichiro Oka:** Writing – review & editing. **Andrew T. Kaczynski:** Investigation, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.dibe.2026.101007.

Data availability

Data will be made available on request.

References

Albadarin, Y., Saqr, M., Pope, N., Tukiainen, M., 2024. A systematic literature review of empirical research on ChatGPT in education. Discover Education 3 (1), 60.

Araos, M., 2023. Democracy underwater: public participation, technical expertise, and climate infrastructure planning in New York city. Theor. Soc. 52 (1), 1–34.

Bedi, S., Liu, Y., Orr-Ewing, L., Dash, D., Koyejo, S., Callahan, A., Fries, J.A., Wornow, M., Swaminathan, A., Lehmann, L.S., 2025. Testing and evaluation of health care applications of large language models: a systematic review. JAMA 333 (4), 319–328.

Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S., 2021. On the dangers of stochastic parrots: can language models be too big?. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp. 610–623.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., 2020. Language models are few-shot learners. Adv. Neural Inf. Process. Syst. 33, 1877–1901.

Cacciatore, S., Mao, S., Nuñez, M.V., Massaro, C., Spadafora, L., Bernardi, M., Perone, F., Sabouret, P., Biondi-Zoccai, G., Banach, M., 2025. Urban health inequities and healthy longevity: traditional and emerging risk factors across the cities and policy implications. Aging Clin. Exp. Res. 37 (1), 143.

Calderon, C., Mutter, A., Westin, M., Butler, A., 2024. Navigating swift and slow planning: planners' balancing act in the design of participatory processes. Eur. Plan. Stud. 32 (2), 390–409.

Christie, C.D., Friedenreich, C.M., Vena, J.E., Doiron, D., McCormack, G.R., 2023. An ecological analysis of walkability and housing affordability in Canada: moderation by city size and neighbourhood property type composition. PLoS One 18 (5), e0285397.

Dannenberg, A.L., Jackson, R.J., Frumkin, H., Schieber, R.A., Pratt, M., Kochtitzky, C., Tilson, H.H., 2003. The impact of community design and land-use choices on public health: a scientific research agenda. Am. J. Publ. Health 93 (9), 1500–1508.

Devon, R., Van de Poel, I., 2004. Design ethics: the social ethics paradigm. Int. J. Eng. Educ. 20 (3), 461–469.

Dheekwal, A., Singh, K., Sharma, A., 2025. GIS-Based modeling of traffic-related air pollution and public health risks: a systematic review with an emphasis on developing urban contexts. Computational Urban Science 5 (1), 65.

Durrant, D., Kossberg, T.N., 2023. Social dramas and planning judgement. Town Plan. Rev. 94 (5), 561–581.

Esmaeilzadeh, P., 2025. Ethical implications of using general-purpose LLMs in clinical settings: a comparative analysis of prompt engineering strategies and their impact on patient safety. BMC Med. Inf. Decis. Making 25 (1), 342.

Frank, L.D., Iroz-Elardo, N., MacLeod, K.E., Hong, A., 2019. Pathways from built environment to health: a conceptual framework linking behavior and exposure-based impacts. J. Transport Health 12, 319–335.

Frumkin, H., 2002. Urban sprawl and public health. Publ. Health Rep. 117 (3), 201.

Fu, X., Wang, R., Li, C., 2024. Can ChatGPT evaluate plans? J. Am. Plann. Assoc. 90 (3), 525–536.

Gallo, R.J., Baiocchi, M., Savage, T.R., Chen, J.H., 2025. Establishing best practices in large language model research: an application to repeat prompting. J. Am. Med. Inform. Assoc. 32 (2), 386–390.

Germani, F., Spitale, G., 2025. Source framing triggers systematic bias in large language models. Sci. Adv. 11 (45) eadz2924.

Goh, E., Gallo, R., Hom, J., Strong, E., Weng, Y., Kerman, H., Cool, J.A., Kanjee, Z., Parsons, A.S., Ahuja, N., 2024. Large language model influence on diagnostic reasoning: a randomized clinical trial. JAMA Netw. Open 7 (10), e2440969.

Griot, M., Hemptinne, C., Vanderdonckt, J., Yuksel, D., 2025. Large language models lack essential metacognition for reliable medical reasoning. Nat. Commun. 16 (1), 642.

Gu, Z., Jia, W., Piccardi, M., Yu, P., 2025. Empowering large language models for automated clinical assessment with generation-augmented retrieval and hierarchical chain-of-thought. Artif. Intell. Med. 162, 103078.

Haltaufderheide, J., Ranisch, R., 2024. The ethics of ChatGPT in medicine and healthcare: a systematic review on Large Language Models (LLMs). npj Digit. Med. 7 (1), 183.

Herkes, F.J., Zouridis, S., 2023. The legitimacy of land use decisions by public authorities in the Netherlands: results from a survey experiment. Land Use Policy 134, 106928.

Kahila, M., Kyttä, M., 2009. SoftGIS as a bridge-builder in collaborative urban planning. In: Planning Support Systems Best Practice and New Methods. Springer, pp. 389–411.

Koohsari, M.J., Kaczynski, A.T., 2025. Make cities more walkable, in the real world and in virtual reality. Nature 646 (8083), 38.

Koohsari, M.J., McCormack, G.R., Li, J., Zhao, J., Long, Y., Lu, Y., Loo, B.P.Y., Oka, K., Kaczynski, A.T., 2025. Can Large Language Models Support Health-Promoting Urban Design?.

Koschinsky, J., Talen, E., 2015. Affordable housing and walkable neighborhoods: a national urban analysis. Cityscape 17 (2), 13–56.

Krapp, A., Barajas, J.M., Wennink, A., 2021. Equity-oriented criteria for project prioritization in regional transportation planning. Transp. Res. Rec. 2675 (9), 182–195.

Kwok, K.O., Huynh, T., Wei, W.I., Wong, S.Y., Riley, S., Tang, A., 2024. Utilizing large language models in infectious disease transmission modelling for public health preparedness. Comput. Struct. Biotechnol. J. 23, 3254–3257.

Leyden, K.M., Slevin, A., Grey, T., Hynes, M., Frisbaek, F., Silke, R., 2017. Public and stakeholder engagement and the built environment: a review. Curr. Environ. Health Rep. 4 (3), 267–277.

Loh, C.G., Kim, R., 2021. Are we planning for equity? Equity goals and recommendations in local comprehensive plans. J. Am. Plann. Assoc. 87 (2), 181–196.

Lucas, H.C., Upperman, J.S., Robinson, J.R., 2024. A systematic review of large language models and their implications in medical education. Med. Educ. 58 (11), 1276–1285.

Matamanda, A.R., Chirisa, I., Mazanhi, P., Toriro, P., 2025. The interface between politics, ethics and urban planning: the case of land and space barons in Harare, Zimbabwe. Environ. Plan. C Politics Space 43 (1), 146–163.

McKay, S., Murray, M., Hui, L.P., 2011. Pitfalls in strategic planning: lessons for legitimacy. Space Polity 15 (2), 107–124.

Messeri, L., Crockett, M.J., 2024. Artificial intelligence and illusions of understanding in scientific research. Nature 627 (8002), 49–58.

Ong, J.C.L., Chang, S.Y.-H., William, W., Butte, A.J., Shah, N.H., Chew, L.S.T., Liu, N., Doshi-Velez, F., Lu, W., Savulescu, J., 2024a. Ethical and regulatory challenges of large language models in medicine. Lancet Digit. Health 6 (6), e428–e432.

Ong, J.C.L., Chang, S.Y.-H., William, W., Butte, A.J., Shah, N.H., Chew, L.S.T., Liu, N., Doshi-Velez, F., Lu, W., Savulescu, J., 2024b. Medical ethics of large language models in medicine. NEJM (7). AI 1AIra2400038.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., 2022. Training language models to follow instructions with human feedback. Adv. Neural Inf. Process. Syst. 35, 27730–27744.

Qiu, P., Wu, C., Liu, S., Fan, Y., Zhao, W., Chen, Z., Gu, H., Peng, C., Zhang, Y., Wang, Y., 2025. Quantifying the reasoning abilities of LLMs on clinical cases. Nat. Commun. 16 (1), 9799.

Rajagopalan, S., Vergara-Martel, A., Zhong, J., Khraishah, H., Kosiborod, M., Neeland, I. J., Dazard, J.-E., Chen, Z., Munzel, T., Brook, R.D., 2024. The urban environment and cardiometabolic health. Circulation 149 (16), 1298–1314.

Raymond, C.M., Nummi, P., von Wirth, T., Poom, A., Ahdekivi, A., Barthel, S., Delmelle, E., Dunkel, E., Fagerholm, N., Grêt-Regamey, A., 2025. Uses, opportunities and risks of artificial intelligence in participatory urban planning. Discover Cities 2 (1), 1–18.

Salazar-Miranda, A., Talen, E., 2025. An AI-based analysis of zoning reforms in US cities. Nature Cities 1–12.

Sanchez, T.W., Brenman, M., Ye, X., 2025. The ethical concerns of artificial intelligence in urban planning. J. Am. Plann. Assoc. 91 (2), 294–307.

Seyedrezaei, M., Becerik-Gerber, B., Awada, M., Contreras, S., Boeing, G., 2023. Equity in the built environment: a systematic review. Build. Environ. 245, 110827.

Smith, M., Hosking, J., Woodward, A., Witten, K., MacMillan, A., Field, A., Baas, P., Mackie, H., 2017. Systematic literature review of built environment effects on physical activity and active Transport–an update and new findings on health equity. Int. J. Behav. Nutr. Phys. Activ. 14 (1), 158.

Stankov, I., Garcia, L.M., Mascolli, M.A., Montes, F., Meisel, J.D., Gouveia, N., Sarmiento, O.L., Rodriguez, D.A., Hammond, R.A., Caiaffa, W.T., 2020. A systematic

review of empirical and simulation studies evaluating the health impact of transportation interventions. Environ. Res. 186, 109519.
Steyvers, M., Tejeda, H., Kumar, A., Belem, C., Karny, S., Hu, X., Mayer, L.W., Smyth, P., 2025. What large language models know and what people think they know. Nat. Mach. Intell. 7 (2), 221–231.
Thirunavukarasu, A.J., Ting, D.S.J., Elangovan, K., Gutierrez, L., Tan, T.F., Ting, D.S.W., 2023. Large language models in medicine. Nat. Med. 29 (8), 1930–1940.
Van Veen, D., Van Uden, C., Blankemeier, L., Delbrouck, J.-B., Aali, A., Bluethgen, C., Pareek, A., Polacin, M., Reis, E.P., Seehofnerová, A., 2024. Adapted large language models can outperform medical experts in clinical text summarization. Nat. Med. 30 (4), 1134–1142.
Watson, V., 2006. Deep difference: diversity, planning and ethics. Plan. Theor. 5 (1), 31–50.
Wong, R.W.M., Loo, B.P.Y., 2025. Using generative AI-Powered sustainable building bots for regulatory intelligence and sustainability assessment of construction: lessons from Hong Kong. J. Urban Technol. 1–24.
Wu, L., Zheng, Z., Qiu, Z., Wang, H., Gu, H., Shen, T., Qin, C., Zhu, C., Zhu, H., Liu, Q., 2024. A survey on large language models for recommendation. World Wide Web 27 (5), 60.
Xia, J., Tong, Y., Long, Y., 2025. Advancements in the application of large language models in urban studies: a systematic review. Cities 165, 106142.
Zack, T., Lehman, E., Suzgun, M., Rodriguez, J.A., Celi, L.A., Gichoya, J., Jurafsky, D., Szolovits, P., Bates, D.W., Abdulnour, R.-E.E., 2024. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. Lancet Digit. Health 6 (1), e12–e22.
Zhang, Y., Lin, Y., Tian, L., Yang, X., 2025a. Leveraging LLM-based multi-agent simulations to boost participatory design education: an experimental exploration in residential area design. Sustain. Cities Soc., 106761
Zhang, Y., Zhao, R., Huang, Z., Long, Y., 2025b. LLMs Capture Urban Science but Oversimplify Complexity. *arXiv preprint arXiv:2505.13803*.
Zheng, Y., Xu, F., Lin, Y., Santi, P., Ratti, C., Wang, Q.R., Li, Y., 2025. Urban planning in the era of large language models. Nat. Comput. Sci. 1–10.