LLM Planning Agents

Exploring the Potential and Challenges of Large Language Model Agents in Urban Design and Planning

Author[1] and Author [2] and Author [3]
[1,2,3]*Affiliation 1.*
[2]*Affiliation 2.*
[1]*Email 1, ORCID 1*
[2]*Email 1, ORCID 2*
[3]*Email 3, ORCID 3*

Abstract. The integration of Large Language Models (LLMs) as planning agents in urban design and planning represents a novel approach to addressing the field's inherent complexity. This study explores their potential and challenges, focusing on their ability to simulate decision-making processes, enhance stakeholder engagement, and provide analytical support. Using an agentic framework, the research evaluates 63 urban development proposals with a specific focus on water management, employing both sequential and nested frameworks. Several LLMs were tested to investigate performance differences across model scales. The findings reveal that while LLM agents exhibit "common sense" and follow planning advice, their reliance on accessible data often results in overly generic outputs, underscoring the need for better data retrieval mechanisms such as Retrieval-Augmented Generation (RAG). Experimental results show nested frameworks outperform sequential ones in reasoning and decision-making, but limitations persist, including biases, limited spatial awareness, and occasional off-topic generation. Addressing these challenges required novel agent architectures and prompt engineering. Smaller models sometimes outperformed larger ones, challenging the assumption that scale guarantees accuracy. Despite these constraints, LLMs demonstrated value in identifying overlooked details and enhancing scenario exploration. This study also advocates for improvements in spatial reasoning, data integration, and framework design.

Keywords. Generative artificial intelligence (GenAI), large language models (LLMs), planning agents, planning application review, urban design and planning

A.B. AUTHOR1, C.D. AUTHOR2 AND E.F. AUTHOR3
FOR FOUR OR MORE AUTHORS USE ‘ET AL.’
- DO NOT ENTER NAMES FOR PEER REVIEW STAGE –

1. Introduction

Urban design and planning is an inherently complex process that requires collective decision-making and collaboration among various stakeholders, as emphasised in Michael Batty's seminal work, "Design as Collective Action" (Batty, 1975). Since the 1970s, intelligent agents have been employed to model these complex processes, enabling a deeper understanding of historical developments and providing predictive insights into future trends (Batty, 2007). The recent advent of Large Language Models (LLMs) and generative models has significantly enhanced the versatility and flexibility of these intelligent agents (Batty, 2024). This advancement allows computational simulations of urban processes to transcend the limitations of the "good old-fashioned AI" (GOFAI) (Haugeland, 1989).

Within this context, contemporary empirical efforts have largely focused on two primary applications: modelling individual activity patterns and motivations (Wang et al., 2024), and simulating resident participation in the planning process (Zhou et al., 2024). Despite these advances, the research field surrounding LLM agents as broad-spectrum planners—encompassing various stakeholders, and even non-human urban entities—is only beginning to take shape. Moreover, the exploration of LLMs' potential and limitations as planning agents remains sparse, representing a significant knowledge gap that impedes applied research in urban planning.

Agentic frameworks, defined as systems where LLMs perform specific tasks and communicate with other agents, have demonstrated their capability to handle complex tasks. These frameworks have shown efficiency in diverse applications, ranging from generating blog posts to drafting entire scientific papers. However, their application to urban planning is still in its infancy. The impressive capabilities of these systems may be constrained when addressing issues like planning or design, where ongoing negotiation and stakeholder involvement are crucial for success. Furthermore, the spatial reasoning abilities of current models are underdeveloped and have yet to be fully integrated into agentic frameworks.

LLMs are increasingly adopted in design and planning, demonstrating enhanced problem-solving abilities through methodologies such as agentic frameworks and chain-of-thought (CoT) reasoning (Wei et al., 2022). However, the application of these advancements to urban and environmental planning is limited. Despite various challenges, planning agents offer substantial research value, particularly when applied at scale across large geographical areas. Assigning an agent to tasks like "reviewing a planning application" may seem impractical at present; however, the outputs of such agents could highlight the clarity and accessibility of planning data, potentially informing policymaking and research. Furthermore, planning agents could play a pivotal role in scenario exploration and improving urban data analysis.

To effectively utilise planning agents as research tools, key questions must be addressed, including how these agents are constructed, the strategies employed for prompting, and the selection of appropriate LLMs. A proposed two-stage experiment seeks to establish foundational guidelines for structuring these agents while addressing issues of model complexity and cost. Future work must evaluate model performance, identify the most suitable frameworks, and establish benchmarks to optimise the deployment of LLMs in urban planning contexts.

2. Objectives of the Study

This study aims to address the existing knowledge gap by proposing an evaluation methodology specifically designed for planning applications. This methodology is then employed to compare various models and agentic frameworks. The findings are used to develop an improved evaluation framework and recommend adjustments to agentic methods, enhancing their long-term performance and practical applicability in urban planning contexts.

3. Methdology

This research employs a methodology centred on the evaluation of urban planning proposals by an LLM-based agentic framework and a human planner on the research team (serving as the ground truth). The proposals are designed as urban development scenarios intended to simulate submissions for preliminary planning application comments. Each proposal is described in textual form and assessed by a planning expert, with a particular focus on water systems and water management.

The agentic framework provides a response in the form of an "approve/reject" decision. To ensure a more nuanced assessment, the agent is also required to provide a rationale explaining its decision. Several LLMs, along with different agentic frameworks, are selected to perform this task. Each model's output is scored by comparing its results against the ground truth established by the expert planner. This scoring system facilitates a systematic evaluation of the models' performance, enabling a comparative analysis of their strengths and limitations.

The following sections elaborate on the methodology, beginning with the selection and scoring of proposals, followed by the details of the models and agentic frameworks employed, and concluding with the approach used for analysing the results.

4. Planning Development Proposals

A series of development proposals were generated for evaluation by the planning agents to analyse and decide upon their approval. These proposals were created by combining a selection of sites, three distinct development projects, and three approaches to be adopted by planners. To enhance the specificity of the task and ensure the research team could trace the root causes and implications of the decisions, the proposals and their assessments focused on water management issues.

Seven sites within the region of West Sussex were selected for the experiment, an area where the research team has familiarity with planning and environmental water challenges. As illustrated in Figure 1 (left), three of these sites are located within the National Landscape of Chichester Harbour, which has a higher protection status and specific water pollution concerns related to the harbour. Two sites are situated in a National Park, where development is likely to face strict restrictions. The remaining two sites lie outside designated or protected areas and are therefore expected to have more relaxed planning frameworks, increasing the likelihood of project approval.

For each site, three projects were tested, each described in a text paragraph outlining

the development of 200 housing units and the corresponding water management strategies. The projects were differentiated by their approach to water management and flood mitigation. Project 1 (high quality) incorporated comprehensive sustainable water management measures, including water reduction, nutrient control, and flood mitigation. Project 2 (moderate quality) adopted some sustainable measures, while Project 3 (low quality) included no such provisions. Consequently, Project 1 was least likely to face rejection, while Project 3 had the highest chances of being rejected.

To determine the assessment value for each proposal, a weighted system was applied, as shown in Figure 1 (right). Each site was assigned an initial score (1-3) reflecting its developmental constraints, with additional scores assigned to the projects (1-3) and approaches (3-1). These were combined into a final weighted score and remapped to the 0-4 scale. Multiple weighting iterations were conducted until the research team considered that the values achieved an acceptable balance of results across sites and approaches. These final assessment values provided a benchmark for comparing the LLM agents' performance (refer to Figure 3 for an ordered list of projects by assessment value).

WEIGHT	
Site	1
Project	1.5
Approach	1

SITE VALUES	
Upwaltham, West Sussex	3
Chichester, West Sussex	2
Graffham, West Sussex	3
Pagham , West Sussex	3
East Wittering , West Sussex	1.5
Bosham , West Sussex	3
Chidham , West Sussex	3

PROJECT VALUES	
Sensitive approach, # 1	1
Neutral approach, # 2	2
Unensitive approach, # 3	3

APPROACH VALUES	
Prioritise nature	3
Neutral	2
Prioritise building	1

Figure 1. Sites and weighting criteria for ground truth

5. Model and Agentic Frameworks Used

An agentic framework refers to a system comprising instances of Large Language Models (LLMs) that are assigned specific functions and structured to interact with one another. Through orchestrated communication and function calling, the collective system can accomplish complex tasks and achieve higher-level reasoning compared to individual LLMs. However, these frameworks have inherent drawbacks, including increased token usage, longer processing times, and the heightened effort required for orchestration and monitoring to prevent potential failures in information flow. This section outlines the agentic frameworks used in this study and the LLMs powering them.

Two types of agentic frameworks were employed in this experiment, differing in how information and messages were passed between agents, which subsequently affected the depth of reasoning and the quality of outcomes. Both frameworks utilised two primary types of agents: research agents and planner agents. Research agents were tasked with gathering constraints or domain-specific information, such as water systems (water expert) or planning and environmental protection (planning expert). The planner agent then considered the proposals alongside the information provided by the research agents to make a final decision (approve/reject).

The primary distinction between the frameworks lies in the flow of information. In the sequential framework, research agents conducted their tasks and passed their findings to the planner without any opportunity for further questioning or interaction. In contrast, the nested framework allowed a more iterative process, where a researcher and a critic could engage in dialogue, reviewing and refining the summarised research before passing it to the planner. This dialogue was recorded, enhancing the depth and context of the assessment.

The sequential framework was implemented using the langchain 0.3.4 library (Das et al., 2024), while the nested framework utilised autogenstudio 0.1.5 and autogen-agentchat 0.2.37 (Wu et al., 2023). Both frameworks employed the Serper service (Serper, 2024) serpapi 0.1.5 to obtain the five most relevant Google search results for predefined queries on water and environmental management. Research agents summarised the information into a concise 200-word paragraph, which was then passed to the planner.

The outcomes of the frameworks were typically presented as conversations between agents. In some instances, these conversations were lengthy, incorporating detailed reasoning alongside the verdict (accept/reject). However, it was challenging to consistently extract single-word decisions due to the tendency of LLMs to include introductory statements or punctuation. To standardise comparisons, responses were passed through a classifier to assign a numerical score (1-4). The one-shot classifier from scikit-llm 1.4.0 tool was used to score responses on a scale from very positive to very negative, corresponding to scores of 0 to 4. This numerical value, termed the assessment value, was compared to the ground truth established by the research team.

Several LLMs of varying sizes were evaluated in this study. These ranged from very small models (Phi-3-mini-128K-Instruct-Q4_0 with 2B, Llama-3.2-3B-Instruct-Q8_0 with 3B), small models (Ministral-8B-Instruct-Q4_0 with 8B), to medium-sized (ChatGPT4o-mini with circa 40B) and large models (ChatGPT4o with circa 500B). Model size, measured in billions of trainable parameters (B), is often assumed to correlate with the quality of outputs. Models under 8B were open-source and executed locally using LMStudio (Studio, 2024), while medium and large models were accessed via API. This range of models enabled a thorough comparison of performance across different scales.

6. Evaluation of Model Performance

As outlined earlier, all 63 proposals were evaluated using the two agentic frameworks powered by the selected LLM models. The assessment value (LLM result mapped onto a scale of 0-4) was compared against the ground truth determined through manual

annotations by the research team. A score of 0 was assigned when the maximum difference occurred between the prediction and the ground truth, while a score of 4 was given when the two were identical. These scores were averaged across sites, projects, and approaches, as well as overall, and are presented in Figure 2.

Additional comparisons were conducted to examine how the assessment values varied across sites, projects, and approaches. Proposals were organised in increasing levels of planning difficulty as defined by the ground truth, starting with those expected to be more easily approved (assessment value = 0) and ending with those deemed highly problematic (assessment value = 4). Ideally, the results produced by the LLMs should align with this order, even if the exact values do not always match. For clarity, values were colour-coded (see Figure 2), with a gradient from green (acceptable) to red (unacceptable) used to visualise the ground truth. LLMs were expected to exhibit a broadly similar colour pattern if they performed well in the exercise.

A similar evaluation was conducted for the average assessment values across sites, projects, and approaches (Figure 2). The aim was to determine whether the overall trends reflected the ground truth, even if individual values deviated.

These evaluations were performed for all tested models and both agentic frameworks, enabling a comprehensive analysis of their relative performance and the alignment of their outputs with the ground truth data.

7. Results

An analysis of the overall scores for both platforms reveals that larger models do not consistently outperform smaller ones within this framework. Notably, the relatively small Llama 3.23B model achieved the highest score in the sequential test (2.2). In general, the variation in scores across models is minor, with most values clustering around 2, the average of the total score range (0–4). As such, these scores alone cannot serve as an absolute measure of model accuracy. Nevertheless, they remain useful for comparing frameworks, particularly as the scores consistently improved from the sequential to the nested framework across all models, except for Phi-3-mini, which was unable to run in the nested configuration.

Sequential framework	
gpt-4o	2.1
gpt-4o-mini	2.0
Ministral-8B	1.9
Llama3.2-3B	2.1
Phi-3-mini	2.0

Nested framework	
gpt-4o	2.2
gpt-4o-mini	2.1
Ministral-8B	2.1
Llama3.2-3B	2.3
Phi-3-mini	-

Figure 2. Overall score of different models

When assessing the models' performance across different sites and projects (Figure 3), the alignment with the ground truth is partial. For example, Phi-3-mini produced predominantly negative recommendations (score: 3), deviating significantly from expectations. The nested framework demonstrated a closer correlation with the ground truth, as evidenced by the alignment of green shades (lower scores) on the left and

orange-red shades (higher scores) on the right. This pattern reinforces the earlier observation of the nested framework's slightly higher overall performance.

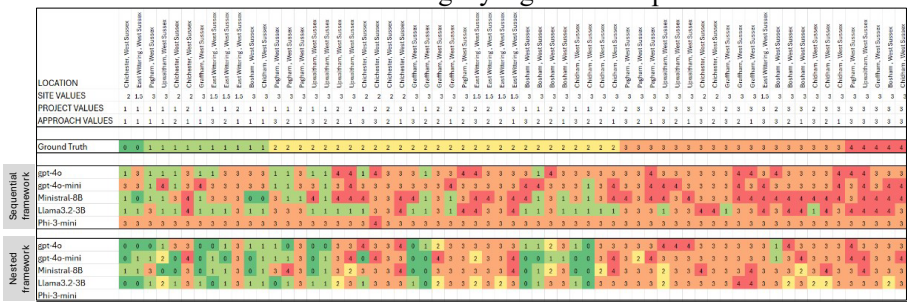

Figure 3. Ordered series of Assessment Values

A deeper examination of site-, project-, and approach-specific values (Figure 4) presents a more complex picture. The correlation between the models and ground truth is weak and, in some cases, counterproductive. For instance, Llama 3.23B, which scored highly overall, failed to align with the ground truth for East Wittering, a site expected to have fewer development constraints (ground truth = 1.6), by assigning overly negative scores (2.7 or 2.4 in the nested framework). Conversely, smaller models like Llama and Ministral demonstrated a variation across projects that more closely matched the ground truth, with lower scores for Project 1 (high quality) and higher scores for Project 3 (low quality). Similarly, models generally assigned higher scores to nature-based approaches (Approach 1) than to more development-oriented ones (Approach 3), indicating some degree of alignment with the input project descriptions and planning objectives.

However, the information extracted from the sites did not add substantial context to distinguish between proposals. This may stem from inadequate research by the agents into site-specific planning constraints. For example, in Chichester, the city outside the National Landscape may have been mistakenly interpreted as part of the broader protected area, leading to excessively conservative decisions.

		Sequential framework					Nested framework				
SITES	**Groudn tr**	gpt-4o	gpt-4o-min	Ministral-8E	Llama3.2-3E	Phi-3-mini	gpt-4o	gpt-4o-min	Ministral-8E	Llama3.2-3E	Phi-3-min
Upwaltham, West Sussex	2.4	2.7	3.3	3.2	1.8	3.0	2.4	2.6	2.0	2.2	-
Chichester, West Sussex	1.8	2.4	2.8	3.1	2.8	3.1	2.9	2.2	2.7	2.2	-
Graffham, West Sussex	2.4	2.8	3.3	2.8	2.0	3.0	2.0	2.1	2.1	2.3	-
Pagham , West Sussex	2.4	2.7	2.4	2.6	3.2	3.0	2.2	2.4	3.1	2.2	-
East Wittering , West Sussex	1.6	2.9	3.1	3.1	2.7	3.0	2.1	2.2	2.4	1.9	-
Bosham , West Sussex	2.4	3.0	3.3	2.8	2.4	3.0	2.1	1.4	1.9	2.0	-
Chidham , West Sussex	2.4	3.1	3.1	3.0	1.8	3.0	2.2	2.3	2.2	2.0	-
PROJECTS	Groudn tr	gpt-4o	gpt-4o-mini	Ministral-8I	Llama3.2-3I	Phi-3-mini	gpt-4o	gpt-4o-min	Ministral-8I	Llama3.2-3	Phi-3-mir
Sensitive approach, # 1	1.4	1.8	2.6	1.9	1.3	3.0	0.8	0.4	1.0	0.7	-
Neutral approach, # 2	2.1	3.3	3.2	3.2	2.3	3.0	2.9	2.9	2.9	2.8	-
Unensitive approach, # 3	3.1	3.3	3.4	3.7	3.6	3.0	3.2	3.2	3.2	2.9	-
APPROACHES	Groudn tr	gpt-4o	gpt-4o-mini	Ministral-8I	Llama3.2-3I	Phi-3-mini	gpt-4o	gpt-4o-min	Ministral-8I	Llama3.2-3	Phi-3-mir
Prioritise nature	2.8	2.9	3.2	3.7	2.5	3.0	2.3	2.1	2.4	2.0	-
Neutral	2.2	2.7	2.8	2.8	2.4	3.0	2.3	2.0	2.3	2.0	-
Prioritise building	1.7	2.8	3.2	2.4	2.2	3.0	2.2	2.4	2.4	2.3	-
AVERAGE	2.2	2.8	3.1	2.9	2.4	3.0	2.3	2.2	2.3	2.1	-

Figure 4. Comparison of Assessment Values

Overall, the models demonstrated a tendency towards conservatism compared to the ground truth. The average ground truth score was 2.1, while most models produced higher averages across both frameworks, with the exception of Llama 3.23B in the nested framework. Notably, the nested framework showed better alignment with the ground truth, partly due to its willingness to approve more projects by avoiding excessive prudence.

These findings highlight the need for enhanced data retrieval and reasoning capabilities in LLMs, particularly for site-specific constraints, to improve alignment with real-world planning assessments.

8. Discussions

During the development of the experiment, the research team identified a significant discrepancy in the results produced by certain LLMs, such as Llama 3.2, when applied to areas like Pegham, which are outside the main conservation zone of Chichester Harbour. A review of the comments generated by the LLMs, combined with further research, revealed the existence of another conservation area that imposed similarly restrictive conditions. This prompted an update to the scoring system, highlighting the importance of allowing the agents to conduct thorough searches before arriving at a final decision.

In terms of programming the agents, the implementation of non-sequential frameworks demanded extensive prompt engineering and control measures. In some cases, the agents engaged in repetitive, off-topic discussions that increased both time and token consumption without yielding relevant outcomes. This phenomenon is not unusual and is commonly observed in natural language generation research and application scenarios. While larger models tended to follow instructions more reliably, these issues were occasionally unavoidable. To address this, a novel agent architecture was tested. Compared to the solution proposed by Jiang et al. (Jiang et al., 2020) , which applied differentiable weights to individual token losses to reduce repetition, this approach allowed planners without advanced AI technical training to use it without needing to modify the underlying algorithms. This ensured feasibility.

Another challenge was the tendency of the agents to "add their opinion" when evaluating proposals. This behaviour introduced a positive bias, complicating the process of extracting clear accept/reject decisions from their assessments. To mitigate this, all models had to be prompted with negative framing to prevent subjective commentary from influencing their final outputs. This, combined with post-hoc methods (Zhang et al., 2023), could potentially further enhance the control over the content generated by LLMs and warrants exploration in future research.

The limited spatial awareness of the models further complicated the analysis. The agents were unable to accurately determine whether a site fell within a protected area based solely on their training data or search results. When a protected area designation appeared during a search, the models often applied this designation indiscriminately, leading them to adopt an overly cautious stance by default. This underscored the need for improved spatial reasoning capabilities and more nuanced processing within agentic frameworks.

9. Conclusions and Further Steps

The results of this experiment demonstrate the potential of agents as tools for planning research and support systems. These agents exhibit a degree of "common sense" and are capable of adhering to planning advice. However, their performance is often limited by the availability of adequate local information, leading to overly generic outputs when such data is insufficient. Addressing this limitation requires ensuring that relevant information is both accessible and easily retrievable. This could involve implementing Retrieval-Augmented Generation (RAG) tailored to specific planning constraints. Notably, the experiment revealed that these models can identify overlooked details, integrating them into the planning framework. Overall, a combination of RAG and search tools emerges as the most effective approach.

The findings also suggest that more complex frameworks enhance the ability of agents to extract contextually relevant information, thereby improving overall performance. These frameworks' capacity to reason and engage in nuanced discussion is a particularly positive feature.

To advance this research, the following improvements are recommended for a more comprehensive study:

- Broader evaluation framework: Future studies should include a wider range of projects representing diverse planning scenarios. With larger datasets, it will be possible to develop methods to better understand the correlation between agent performance and evaluation metrics, such as paired t-tests or Pearson correlation coefficients.
- Enhanced data retrieval: Implement RAG or other data retrieval mechanisms specifically designed for urban planning, providing accurate spatial information and addressing planning constraints more effectively.
- Integration of spatial and visual intelligence: Introducing models capable of spatial reasoning and visual analysis could improve the agents' situational awareness and overall performance.
- Testing advanced agentic frameworks: Further exploration of more sophisticated frameworks may enhance agents' reasoning and contextual understanding.

Finally, this research could evolve towards a stronger generative approach, where agents not only analyse and comment on projects but also generate proposals or suggest improvements. Such advancements would further establish the role of intelligent agents as transformative tools in urban and environmental planning.

References

Batty, M. (1975) Design as Collective Action. *Environment and Planning B: Planning and Design*, 2, 151-176. https://doi.org/10.1068/b020151

Batty, M. (2007) *Cities and Complexity: Understanding Cities with Cellular Automata, Agent-Based Models, and Fractals*, Cambridge, MA, MIT Press.

Batty, M. (2024) AI and Design. *Environment and Planning B: Urban Analytics and City Science*, 51, 799-802. https://doi.org/10.1177/23998083241236619

Das, D., Rath, R. L., Singh, T., Mishra, S., Malik, V., Sobti, R., & Brahma, B. (2024). Llm-Based Custom Chatbot Using Langchain. *International Conference on Innovative Computing And Communication* https://doi.org/ 10.1007/978-981-97-3588-4_22

Haugeland, J. (1989) *Artificial Intelligence: The Very Idea*, Cambridge, MA, MIT Press.

Jiang, S., Wolf, T., Monz, C. & De Rijke, M. (2020) *TLDR: Token Loss Dynamic Reweighting for Reducing Repetitive Utterance Generation*. arXiv preprint arXiv:2003.11963.

Serper. (2024). *Serper: The World's Fastest & Cheapest Google Search API*. https://serper.dev/

Studio, L. (2024) *LM Studio*. LM Studio.

Wang, J., Jiang, R., Yang, C., Wu, Z., Onizuka, M., Shibasaki, R. & Xiao, C. (2024) *Large Language Models as Urban Residents: An LLM Agent Framework for Personal Mobility Generation*. arXiv preprint arXiv:2402.14744.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V. & Zhou, D. (2022) Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in Neural Information Processing Systems*, 35, 24824-24837. https://doi.org/ 2201.11903

Wu, Q., Bansal, G., Zhang, J., Wu, Y., Zhang, S., Zhu, E., Li, B., Jiang, L., Zhang, X., & Wang, C. (2023). *Autogen: Enabling Next-Gen Llm Applications Via Multi-Agent Conversation Framework*. arXiv preprint arXiv:2308.08155.

Zhang, M., Sokolov, A., Cai, W. & Chen, S.-Q. (2023) *Joint Repetition Suppression and Content Moderation of Large Language Models*. arXiv preprint arXiv:2304.10611.

Zhou, Z., Lin, Y., Jin, D. & Li, Y. (2024) *Large Language Model for Participatory Urban Planning*. arXiv preprint arXiv:2402.17161.

.