



Generative Agents for High-Fidelity Simulation of Community-Scale Mobility and Energy Behavior

Yongjian Chen, Zhifeng Yang, and Shiqi(Shawn) Ou South China University of Technology

Citation: Chen, Y., Yang, Z., and Ou, S.(S.) "Generative Agents for High-Fidelity Simulation of Community-Scale Mobility and Energy Behavior," SAE Technical Paper 2026-01-0465, 2026, doi:10.4271/2026-01-0465.

Received: 19 Oct 2025

Revised: 16 Dec 2025

Accepted: 26 Jan 2026

Abstract

The transition to sustainable mobility and energy systems represents a complex socio-technical challenge, with the success of new technologies and policies critically dependent on their interaction with human behavior. Traditional models frequently struggle to capture the nuanced, heterogeneous, and adaptive characteristics of individual decision-making in mobility choices and energy usage, thereby introducing significant uncertainties into system design and policy evaluation. This paper presents a novel paradigm to bridge this gap: the Hierarchical Generative Agent-based Simulation Framework (HGA-Sim). The framework's core innovations are twofold: 1) It utilizes Large Language Models to generate agents endowed with intrinsic personality traits autonomously, enabling a realistic simulation of diverse,

human-like responses to environmental stimuli and personal experiences. 2) It employs a hierarchical "Archetype-Individual" architecture, rendering large-scale community simulations computationally feasible. Validated through a case study of a 495-agent community, the HGA-Sim framework accurately reproduces aggregate mobility and energy consumption patterns, including critical peak loads and temporal dynamics. It demonstrates remarkable fidelity to real-world data with a Root Mean Square Error of 0.1983. By establishing a human-in-the-loop virtual testing environment, this research provides a foundational tool for evaluating the real-world viability of sustainable mobility designs, assessing the potential socioeconomic impacts of new transportation and energy policies, and mitigating risks associated with investments in future sustainable infrastructure.

Introduction

In the context of escalating global climate change and environmental pollution, the building sector, a major contributor to energy consumption and carbon emissions, faces unprecedented pressure to reduce its environmental footprint. Statistically, buildings account for 34% of global energy consumption and a staggering 37% of energy-related carbon dioxide emissions [1]. As a collection of architectural facilities, campus energy management and carbon emission control have thus become pivotal for achieving sustainable development goals, with electricity consumption widely identified as a primary source of emissions. Consequently, accurately understanding and effectively improving the energy efficiency of campus buildings has emerged as an urgent global imperative.

Extant research consistently demonstrates that occupant behavior patterns and their spatial distribution are critical variables influencing building energy consumption, particularly electricity usage. For instance, studies by Socolow et al. revealed that up to 71% of the variations in energy demand in residential buildings could be attributed to differences in occupant behavior [2]. A complex,

interactive relationship exists between occupants and energy systems, with the stochastic, diverse, and adaptive nature of human behavior being the largest source of uncertainty in predicting building energy consumption. Therefore, high-fidelity simulation of occupant behavior is crucial not only for profoundly uncovering the root causes of the "performance gap" between designed and actual operational performance in campus environments but also for providing essential decision support for formulating scientifically sound energy management strategies and carbon reduction pathways [3,4].

To investigate the relationship between occupant behavior and energy consumption, various modeling approaches have been developed within academia, primarily categorized into physical models, stochastic models, data mining models, and traditional Agent-Based Models (ABM) [5]. Physical models, such as EnergyPlus, simulate energy consumption by defining static occupant schedules and fixed equipment parameters; however, they struggle to capture the nonlinear energy-use patterns and the dynamic randomness of occupant behavior in the real world. Stochastic models, such as Markov chains, model occupant presence as a random process but often

oversimplify the underlying decision-making logic of behavior. Data-driven methods aim to build mappings between occupancy and energy consumption by mining historical data. Nevertheless, these approaches heavily rely on large-scale, high-quality datasets, and acquiring sensitive private data is challenging, significantly limiting the generalizability of such models. While traditional ABMs can, to some extent, simulate adaptive individual behaviors (e.g., adjusting attire, switching appliances on/off), their behavioral logic still depends on predefined rules and heuristic algorithms, making it difficult to fully replicate the deep, nuanced, and personalized cognitive processes involved in human decision-making.

In recent years, the integration of generative artificial intelligence with agent-based modeling—termed Generative Agent-Based Model (GABM)—has offered a novel paradigm for high-fidelity simulation of occupant behavior [6]. The core of GABM lies in leveraging the extensive knowledge base and natural language reasoning capabilities of pre-trained Large Language Models (LLMs) to drive agent decision-making processes [7]. Trained in vast amounts of textual data, LLMs can exhibit emergent common-sense reasoning and situational decision-making abilities that are highly akin to those of humans. Park et al. successfully demonstrated the capacity of 25 generative agents in a virtual town to interact and generate credible social dynamics autonomously [8]. Subsequently, GABM has been successfully applied across various domains, including epidemic spread [9], voting behavior [10], and macroeconomic simulation [11]. Its defining characteristic is that agent behaviors are autonomously generated by LLMs in specific contexts, rather than relying on hard-coded, predefined rules. This approach allows the entire system to dynamically evolve with agent decisions, forming a complete feedback loop. Almashor et al. have made preliminary attempts to simulate household energy consumption behavior using GABM, achieving energy consumption patterns consistent with real data [12]. However, applying GABM to simulate energy consumption behavior in large-scale, complex campus scenarios remains in its nascent stages. This is primarily constrained by two major challenges: first, configuring user personas for large-scale agents to reflect personality traits and behavioral diversity is exceedingly cumbersome and labor-intensive; second, as the scale of scenarios and agents drastically increases, the computational resource demands (especially memory) and the invocation costs of LLMs APIs grow exponentially, severely limiting the model's scalability and economic viability.

To address the challenges mentioned above, this study proposes the Hierarchical Generative Agent-based Simulation Framework (HGA-Sim) for campus energy consumption behavior, using a university campus as an empirical case study. To address the complexity of modeling individual behavior, this study introduces a multidimensional user persona framework that integrates personality traits. This framework leverages the reasoning capabilities of large language models to autonomously generate agents with differentiated decision-making logic, thereby overcoming the simplified assumptions of

traditional models regarding the randomness of human behavior. To circumvent computational bottlenecks in large-scale simulations, we innovatively designed a hierarchical "Archetype-Individual" decision-making architecture. By clustering agent archetypes, individual decisions are generalized into a limited set of typical paradigms. This significantly reduces LLM invocation costs while preserving individual behavioral diversity through probabilistic sampling, effectively balancing scale and economic viability in mobility and energy consumption behavior simulations. The main contributions of this study can be summarized as follows:

1. **Automated Generation of Personality-Based User Personas:** By integrating psychological personality traits (e.g., the Big Five personality traits) into GABM and employing carefully designed Prompt Engineering, we achieved automated and personalized initial settings for agent user personas, significantly reducing the complexity and subjectivity of manual configuration in large-scale simulations.
2. **Scalable Hierarchical Decision Architecture:** A novel "Archetype-Individual" dual-layer decision-making architecture is proposed. This architecture generates representative "archetype agents" through clustering algorithms to handle common decisions, and then combines probabilistic distributions to guide the specific actions of "individual agents." This approach enhances the computational efficiency and economic viability of large-scale simulations while ensuring behavioral diversity.
3. **Multi-Scenario Large-Scale Campus Energy Consumption Simulation and Validation:** This study marks the first application of GABM to campus-level energy consumption simulation. By constructing a campus simulation environment with 495 agents, we simulated their mobility and energy consumption behaviors. The effectiveness and practical value of the proposed framework were validated by comparing the simulation results with real-world energy consumption patterns.

The remainder of this paper is organized as follows: Section 2 elaborates on the HGA-Sim's framework and methodology. Section 3 presents the specific settings of the case study, simulation results, and their analysis and implications. Finally, Sections 4 and 5 conclude the paper and outline future research directions.

Methodology

The HGA-Sim consists of two main components: the generative agent architecture and the "Archetype-Individual" architecture. Each generative agent is equipped with functions encompassing memory, planning, reflection, and decision-making. Interactions with the campus

sandbox environment are achieved by parsing the natural language outputs generated during the agents' decision-making process, thereby producing high-fidelity energy consumption data.

Specifically, the system first constructs a large population of individual agents with diverse roles and distinct personality traits. Subsequently, K-means clustering is employed to identify different "archetype agents." This approach aims to reduce the computational load associated with LLM calls, thus enabling large-scale simulations. Various archetype agents then invoke LLMs based on perceived information, memories, and plans to generate behavioral patterns. These archetype behavioral patterns subsequently serve as a reference for individual agent decision-making. After stochastic sampling and behavior path matching, these patterns are adopted as the behavioral trajectories of particular agents. Within the campus sandbox environment, individual agents can interact with different "tile" objects. Interactions with electrical appliances are recorded by the campus sandbox system, which then aggregates the data to derive the overall energy consumption of the campus sandbox environment.

Generative Agent Architecture

A generative agent typically comprises four modules: memory, reflection, planning, and action. The memory, reflection, and planning modules for archetype agents are designed based on the architecture proposed by Park et al. [8] as shown in Figure 1 and will be introduced in this subsection. In the hierarchical "Archetype-Individual" framework proposed in this paper, the agent's behavioral decision-making module is a core innovation and will be presented separately in the next subsection.

Memory and Retrieval Module To enable generative agents to realistically simulate human behavior, the system must process complex experiential data that extends far beyond the typical capacity limits of traditional prompts. In this context, we refer to all experiences an agent has undergone as "memories." Programmatically, an agent's memory stream is structured as a chronological series of objects. Each object encapsulates the event's content, described in natural language, along with a creation timestamp and the most recent access timestamp. This stream comprehensively records diverse events directly perceived by the agent, encompassing both its own actions and observed external occurrences (e.g., "the overhead light was turned on"). Directly feeding

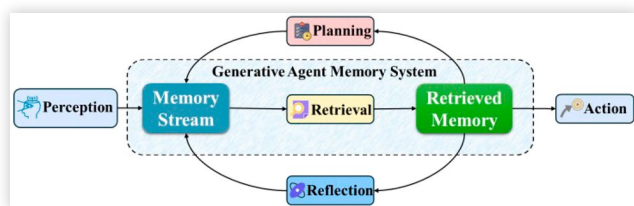
all memories into a prompt has obvious drawbacks: on the one hand, a complete memory stream can dilute the model's focus, and on the other hand, the context window of current models cannot accommodate such a large amount of information. Taking David, an agent in the role of a student, as an example: when asked, "Where are you planning to go next?", simply stuffing all of David's experiences into the prompt might yield a generalized and meaningless answer, such as "want to go to the library to find a quiet place to study." However, through the memory stream mechanism designed in this system, more contextually relevant memory fragments can be intelligently extracted. For instance, when David is facing a paper deadline, a more informative and targeted response can be generated: "Plan to work late in the office."

To enable efficient information extraction from a vast memory stream, a context-based intelligent retrieval module has been implemented. This module comprehensively considers three core dimensions to filter the most relevant subset of memories for the current situation: the Recency (S_{rc}), the Importance (S_{rc}), and the Relevance (S_{rl}). First, recency assigns higher weights to recent memories via a time-based exponential decay function, ensuring the timeliness of events. Second, importance aims to distinguish critical memories (e.g., "attended an important meeting") from daily trivialities (e.g., "had breakfast"); its score is innovatively generated directly by the large language model based on the event's emotional and narrative weight. Finally, relevance is quantified by calculating the cosine similarity between the memory's embedding vector and the current query vector, ensuring semantic proximity. To generate the final ranking, the system normalizes the scores from these three dimensions. It calculates a comprehensive score via a weighted sum, as shown in Equation (1), ultimately injecting the highest-ranked memory content into the LLM's context window as critical input for its subsequent thinking and decision-making. All weights referenced in [8] are set to 1.

$$Score = w_{rc} \times S_{rc} + w_{ip} \times S_{ip} + w_{rl} \times S_{rl} \quad (1)$$

Reflection Module To simulate human meta-cognitive abilities, a "reflection" module has been constructed. Its core function is to drive agents to abstract and summarize disparate experiences periodically. This reflection mechanism is activated when the cumulative importance score of an agent's recent experiences exceeds a predefined threshold of 150. The entire process begins with an introspective questioning phase: by inputting recent memory fragments into the large language model, it is prompted to generate several deep and broad exploratory questions. These questions are subsequently used as keywords for retrieval, gathering all relevant historical memories as analytical material. Based on this, the model is guided to perform inductive reasoning, extract new insights or conclusions, and trace back to link these conclusions to their original supporting evidence. All abstract thoughts generated through reflection are ultimately stored in a tree-like hierarchical structure, with raw observational

FIGURE 1 Generative Agent Framework



data at the bottom and progressively abstract concepts at higher levels, thus forming a dynamically evolving cognitive network.

Planning Module To ensure temporal coherence and logical consistency in agent behavior, this system has implemented a planning mechanism. Solely relying on the immediate responses of LLMs can lead to distortions in behavioral sequences, such as repetitive, illogical behaviors (e.g., repeatedly eating meals) within a short period. Planning provides a crucial framework for agents to maintain behavioral consistency by pre-setting a series of action sequences with temporal and spatial attributes. A plan typically includes key elements such as the action's location, start time, and expected duration. This planning information is stored as a special type of memory within the memory stream, forming a reference basis for agent decision-making alongside observations and reflections. Furthermore, this mechanism supports dynamic adjustment, enabling agents to modify established plans during execution in response to changes in context.

In the plan generation phase, the system first uses the LLM to generate a daily coarse-grained plan for the archetype agent, including wake-up time allocation, 24-hour activity arrangements, and corresponding spatial area mapping. The archetype agent uses the memory retrieval mechanism to gather relevant historical experiences, combining contextual information like the current time, special events, and daily habits to create a diverse schedule. Subsequently, the system further decomposes the coarse-grained hourly plans into fine-grained sub-task sequences, where each sub-task includes specific activity descriptions, durations, and spatial location information. Through this step-by-step decomposition strategy, which progresses from macroscopic schedules to microscopic tasks, this system ensures not only the long-term consistency of agent behavior with its goals but also guarantees the rationality and detailed realism of its actions within short time scales.

Archetype and Individual Agent Architecture

This paper aims to construct a hierarchical "Archetype-Individual" architecture to effectively reduce the computational overhead of community-level simulations and achieve a harmonious blend of coordinated unity and heterogeneity in agent behavior.

Agent Configuration and Automated Generation To address the time-consuming, labor-intensive challenges of configuring agent simulation systems, this study proposes an automated method for generating user personas based on personality traits. This method combines random sampling with an LLM to automatically configure agents' user persona attributes, ensuring both diversity in user personas and avoiding the need for cumbersome manual settings.

In this system, each agent includes initial attributes such as basic, perceptual, and location information. Details of their data dictionary can be found in [Table 1](#) and [Table 2](#). All agents' initial information is stored as JSON files and embedded in the prompt when the LLM is invoked. Among these, plans, current status descriptions, and spatial perception will dynamically update as the simulation progresses. Archetype agents possess all the aforementioned attributes, while individual agents only include Age, Role, Spatial Cognitive Tree (Spatial), and Personality attributes.

[Figure 2](#) illustrates the flowchart of our proposed method. First, the total number of agents in the community, role proportions, personality traits (e.g., Big Five personality traits), and spatial preference parameters for each role are defined. Subsequently, age, name, personality traits, and spatial tree are randomly generated for all agents. Age is randomly sampled within the role's predefined numerical range; personality traits are generated through normal distribution sampling; the spatial tree is initialized with a tree-like structure by selecting buildings, areas, and equipment based on spatial preference parameters; and the LLM generates names. By following the above steps, a batch of individual agents

TABLE 1 Basic attributes of agents

Attribute	Type	Description	Example Value / Range
Id	Integer	Unique identifier for the agent	1-99
Name	String	Agent's name	"Alice", "Bob", "Charlie", ...
Coord	Integer Array	[x, y] coordinates on a 2D map	[82, 9], [10, 8]
Currently	String	Description of current status	"Violet is resting at the Dormitory."
Age	Integer	Agent's age	18-65
Role	String	Role description	"faulty"
Lifestyle	String	Description of lifestyle	"Gets up early, goes to work, has meals in the Cafeteria..."
Daily_plan	String	Brief description of daily plan	"Rest in the dormitory."
Spatial	Object	Agent's spatial cognitive tree	"tree": {"the Campus": {...}}

TABLE 2 Agent Personality Traits

Attribute	Type	Description	Range
extraversion	Integer	Extraversion	1-9
openness	Integer	Openness	1-9
conscientiousness	Integer	Conscientiousness	1-9
agreeableness	Integer	Agreeableness	1-9
neuroticism	Integer	Neuroticism	1-9

can be automatically generated. To select representative archetypes from the community-scale agent population, this study performs K-means clustering on individual agents' personality traits, selecting the agents closest to each cluster center as archetype agents for different roles. Finally, the configured agent parameters are used as prompts to the LLM to generate the archetype agents' Lifestyle, Daily Plan, and Currently attributes. These attributes provide crucial contextual information for subsequent agent prompts, enriching the character features of agents.

Agent Mobility and Energy-Use Behavior In the archetype-individual dual-layer decision-making framework proposed in this paper, the archetype agent perceives its surrounding environment at each time step and stores the perceived information in memory. The LLM then integrates the agent's current attributes and perceptual information to output behavioral decisions at each moment. For example, at 8:00 a.m., a student agent typically moves to the teaching building, demonstrating mobility. Upon arriving in the classroom, the student agent may turn on the lights, reflecting energy-use behavior. These decisions are generated by the LLM rather than being pre-defined, which is one of the key distinctions from traditional ABM. Detailed prompt examples and agents' decision responses are provided in the Appendix.

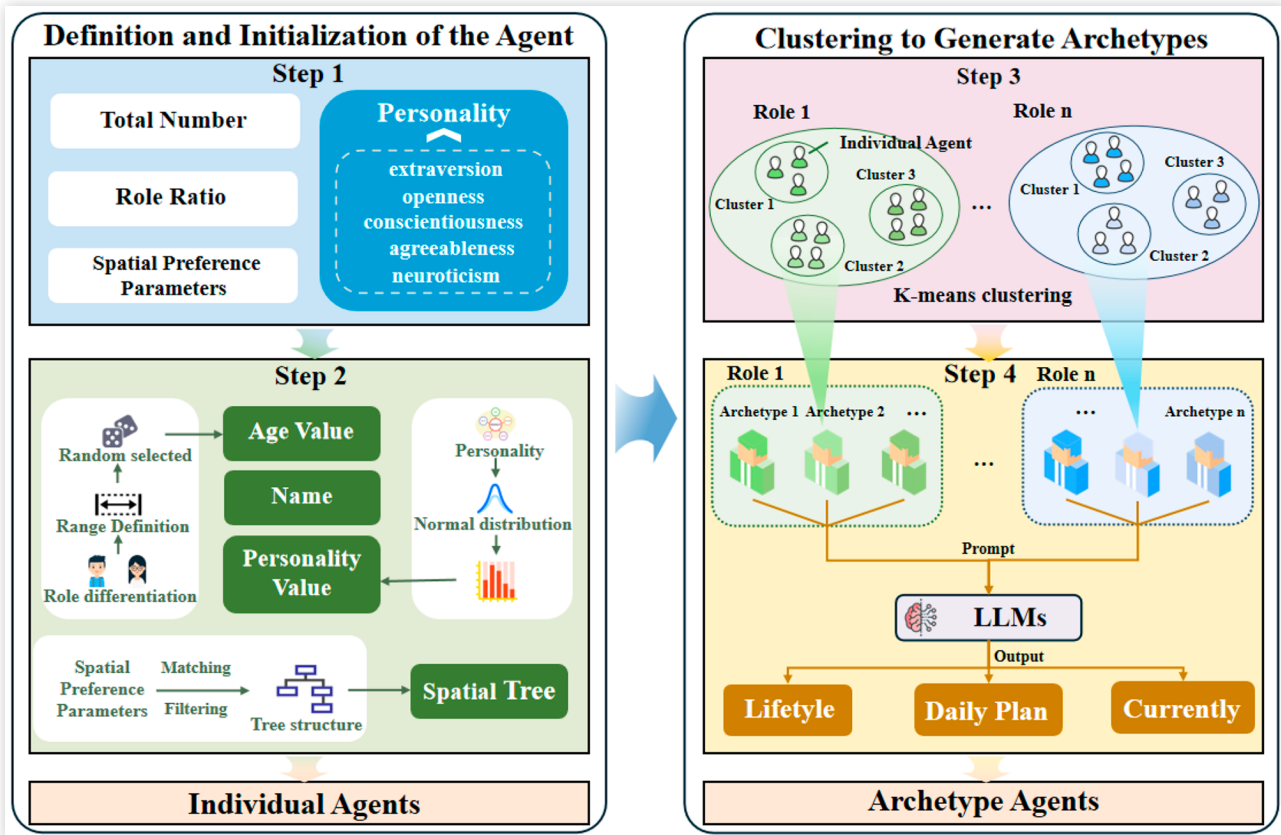
The archetype-individual hierarchy is a layered architecture in which behavioral decisions are initiated by the archetype and then driven top-down to the individual level. While this approach theoretically reduces computational overhead, it faces the challenge of preserving behavioral differentiation among individual agents. In the hierarchical architecture designed in this study, the archetype influences individuals through three parameters: time, space, and behavior, as illustrated in Figure 2.

Regarding the time parameter, there is a probabilistic distribution in the action time difference between the archetype and its associated individuals. At the start of the day (wake-up time), an individual's wake-up time follows a normal distribution with the mean set as the archetype's wake-up time plus n simulation time steps and a variance of m time steps. This design ensures that the agents' action times are smooth and more realistic.

Regarding the space parameter, after the archetype determines a behavioral decision, its behavioral path serves as a reference for the individual agents. Individuals match and retrieve the most similar behavioral path from their perceptual tree as their decision basis. The perceptual tree structure consists of four layers: world, building, area, and object, as shown in the lower part of Figure 2.

On the behavior parameter level, not all individuals strictly follow the archetype's behavioral path. A certain proportion of agents make decisions based on their

FIGURE 2 Flowchart of automated agent generation.



historical behavior preferences. If no historical behavior exists, they randomly sample a behavioral path from the perceptual tree.

Through these three dimensions, the archetype-individual architecture achieves hierarchical decision-making, forming an LLM-driven top-down interaction mode that balances computational cost and behavioral heterogeneity.

Results

This study applies the proposed HGA-Sim to a manually designed campus community. The primary aim is to validate its effectiveness in simulating large-scale occupant behavior and predicting energy consumption. The experiments were conducted on a PC equipped with an AMD Ryzen 7 9700X CPU and an NVIDIA RTX 4070 Super GPU. Ollama [13] was used to deploy Qwen2.5-7b locally [14] as the large language model. The simulation generated data on occupant mobility and building energy consumption within the campus community at a 5-minute resolution.

Campus Community Case Study Information

To contextualize the subsequent analysis and findings, this section establishes a detailed profile of the campus community under investigation. This profile encompasses four key aspects: the primary energy-consuming equipment, the methodology for calculating energy consumption, the design of the campus environment, and the relevant occupant information.

Energy-Using Equipment and Energy Consumption Calculation In modeling the energy-using behavior of occupant agents, this research categorizes residential loads into two typical patterns: instantaneous loads and base loads. Instantaneous loads are power-consuming devices directly controlled by occupants that operate at their rated power for specific durations, such as lighting systems, personal computers, and projectors. The activation and deactivation of these loads exhibit significant temporal randomness, entirely dependent on the actual usage demands of the occupants. The cumulative energy consumption C for a total of m devices is calculated using Equation (2):

$$C = \sum_{k=1}^m (P_k \times T_k) \quad (2)$$

Here, P_k denotes the power of the k -th device, and T_k represents the cumulative operation time of the k -th device, which the occupant agents determine.

Environment and Occupant Information In this case study, four main buildings were defined: the Dormitory, the Cafeteria, the Institute Building, and the Teaching

Building, as illustrated in Figure 3. Each building area has clearly defined coordinate ranges to ensure a rational spatial layout. A grid-based distribution algorithm was employed to subdivide the building interiors into multiple room areas, with connecting corridors established between rooms to form a comprehensive circulation network. The Cafeteria is divided into dining and kitchen areas; the Dormitory into multiple residential rooms; the Institute Building into offices and meeting rooms; and the Teaching Building into multiple classrooms, computer labs, and a Lecture Theater. The number and layout of equipment within the rooms were designed to reflect realistic conditions.

Based on the size of the campus community, a total of 495 occupants were configured, comprising four roles: undergraduates, postgraduates, faculty, and staff, with numbers of 198, 149, 99, and 49, respectively. All occupants were initialized according to the proposed HGA-Sim methodology. During the clustering phase, after determining the optimal number of clusters using silhouette coefficients, the four roles were clustered into 10, 9, 7, and 5 archetypes, respectively. Finally, the LLM is utilized to complete the basic attribute settings for these archetypes.

Occupant Behavior and Energy Consumption Results

Occupant Behavior Characteristics The proposed HGA-Sim simulates the evolution of a 24-hour workday (from 0:00 to 0:00 the following day). The statistical changes in occupant numbers across buildings in a single simulation run are shown in Figure 4. This figure visually demonstrates the trends and directions of pedestrian flow within the campus over a simulated day. To better analyze population mobility characteristics, different background colors indicate approximate mealtimes: bright yellow for breakfast (8:00-9:00), light orange for lunch (12:00-13:00), and blue-purple for dinner (18:00-19:00).

As observed from the figure, in the early hours before breakfast, all occupants are resting in the Dormitory. As breakfast time approaches, occupants begin to leave the Dormitory in large numbers, heading towards the other three buildings. During the 9:00-12:00 period, when occupants are working and studying, the Teaching Building and Institute Building experience peak occupancy. During lunch, the Cafeteria sees its peak attendance. In the afternoon, a large portion of the population returns to the Teaching Building. The Cafeteria again experiences a peak in numbers during dinner time. As night approaches, occupants gradually return to the Dormitory to rest. Overall, the daily population flow characteristics closely align with those of a real campus, indicating that the proposed method effectively captures occupants' mobility behavior within a campus community.

Energy Consumption Characteristics Through 5 repeated experiments, with the time resolution aggregated to 1 hour and then normalized, the energy

FIGURE 3 The hierarchical "Archetype-Individual" dual-layer decision framework.

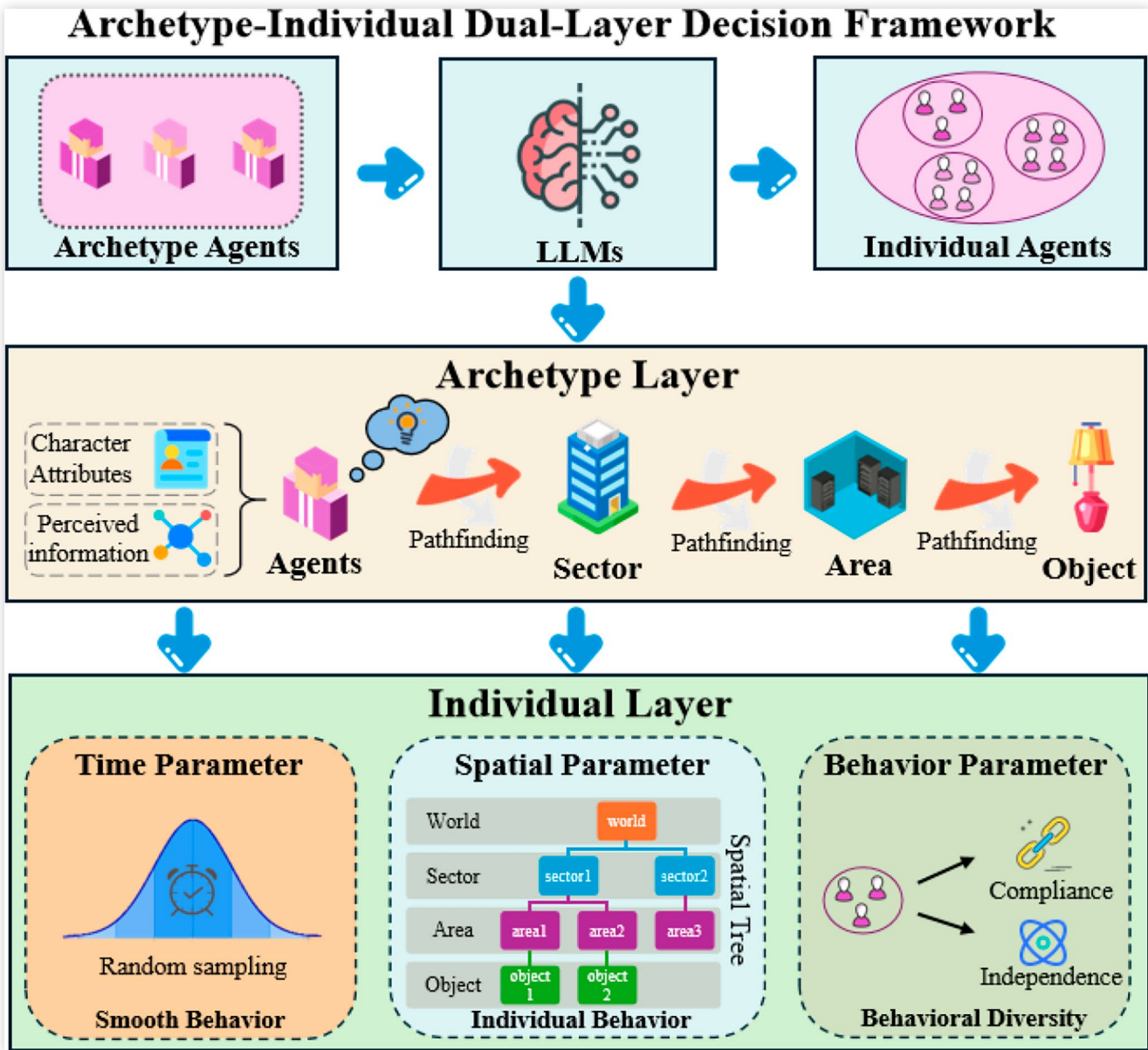


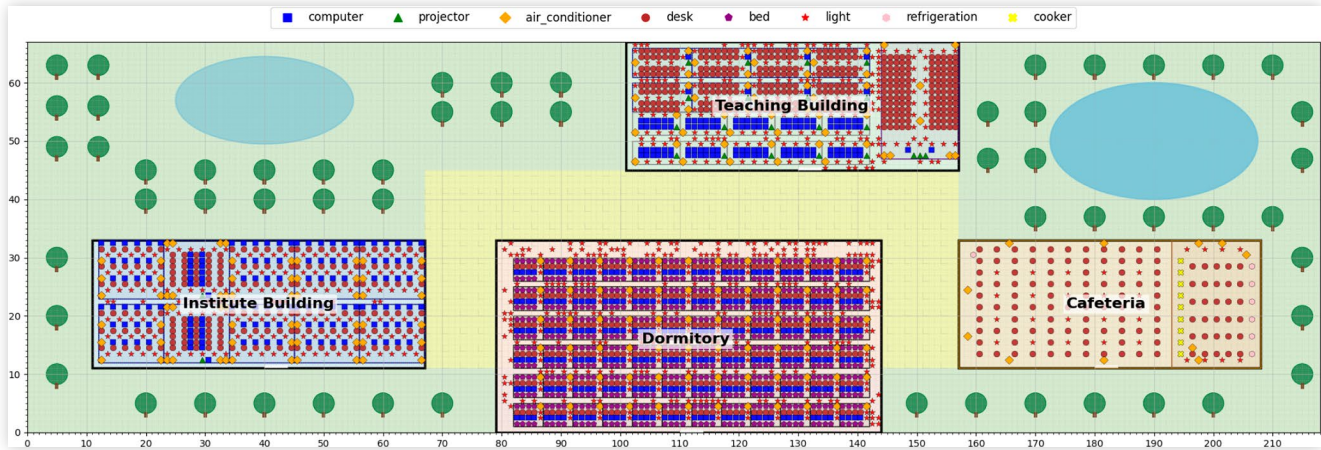
TABLE 3 Device power settings.

Mode	Device Type	Power
Instantaneous Load	Air Conditioner	1000W
	Electric Light	370W
	Projector	150W
	Computer	100W
Base Load	Refrigerator	100W
	Digital Clock	3W
	Router	5W

consumption results for each building are presented in Figure 6. The figure plots three lines representing the maximum, mean, and minimum values at each time point across the repeated experiments, while the box plots illustrate the distribution of the data.

For the Cafeteria, a distinct bimodal characteristic is observed around lunch and dinner times. The Dormitory experiences an energy consumption peak around 8:00 AM, primarily due to occupants' energy-using behaviors upon waking. The Institute Building and Teaching Building exhibit a trimodal pattern in the morning, afternoon, and evening, reflecting energy demand during three distinct work and study periods.

Comparing the box plot characteristics, it is evident that the energy consumption fluctuations for the Cafeteria and Institute Building are larger than those for the other two buildings. Two main reasons are hypothesized for this observation: 1) There is some variation in the generative agents' decisions regarding mealtimes across different simulation runs. 2) The functional definition of the Institute Building might be less clear than that of the other three buildings, leading to variations in the agents'

FIGURE 4 Sandbox map for the campus community case.

interest in visiting this location across different simulation iterations.

To further validate the method's effectiveness, the average energy consumption values for the four buildings were compared with actual campus building energy consumption data. The real energy consumption data was obtained from Shandong Jianzhu University in China [15], where remote data collectors monitor energy usage across various energy-consuming units (buildings, areas, rooms) on campus. The comparison results, shown in Figure 7, indicate that although the peak times of the simulated and real values do not perfectly coincide, they exhibit similar energy consumption trend characteristics. The calculated Root Mean Square Error (RMSE) is 0.1983. This result confirms that the proposed HGA-Sim can effectively generate energy consumption patterns that closely approximate real-world scenarios.

Discussion

While the experimental results in Section 4 have demonstrated the high fidelity of HGA-Sim in reproducing

mobility and energy patterns (RMSE = 0.1983), the value of this framework extends beyond mere accuracy. In this section, we discuss the distinct advantages of HGA-Sim over traditional modeling paradigms and provide a quantitative analysis of the computational efficiency gains achieved by our hierarchical architecture.

Unlike traditional ABMs that rely on static schedules or predefined heuristic rules, HGA-Sim represents a paradigm shift towards generative simulation. Table 4 summarizes the fundamental distinctions between the proposed framework and traditional methods. By leveraging the autonomous reasoning capabilities of Large Language Models, HGA-Sim overcomes the rigidity of "If-Then" logic, enabling the simulation of nuanced, non-linear human decision-making processes. Furthermore, automated user persona generation significantly reduces configuration complexity. At the same time, the model's zero-shot reasoning capability offers superior adaptability to undefined scenarios—advantages that are critical for future policy testing and emergency response planning.

To quantitatively validate the improvement in computational efficiency brought by the proposed architecture, a token usage analysis was conducted by parsing the logs returned by the Ollama API. The statistical results indicate

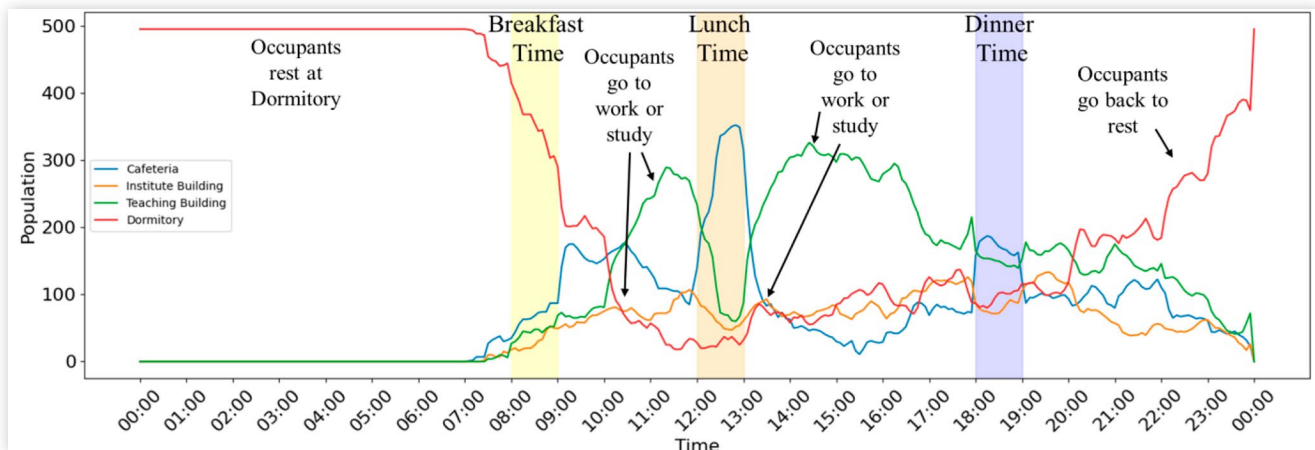
FIGURE 5 Time series of occupants counts in four campus buildings (Cafeteria, Institute Building, Teaching Building, and Dormitory).

FIGURE 6 Energy consumption variation range for the four campus buildings (Cafeteria, Institute Building, Teaching Building, and Dormitory).

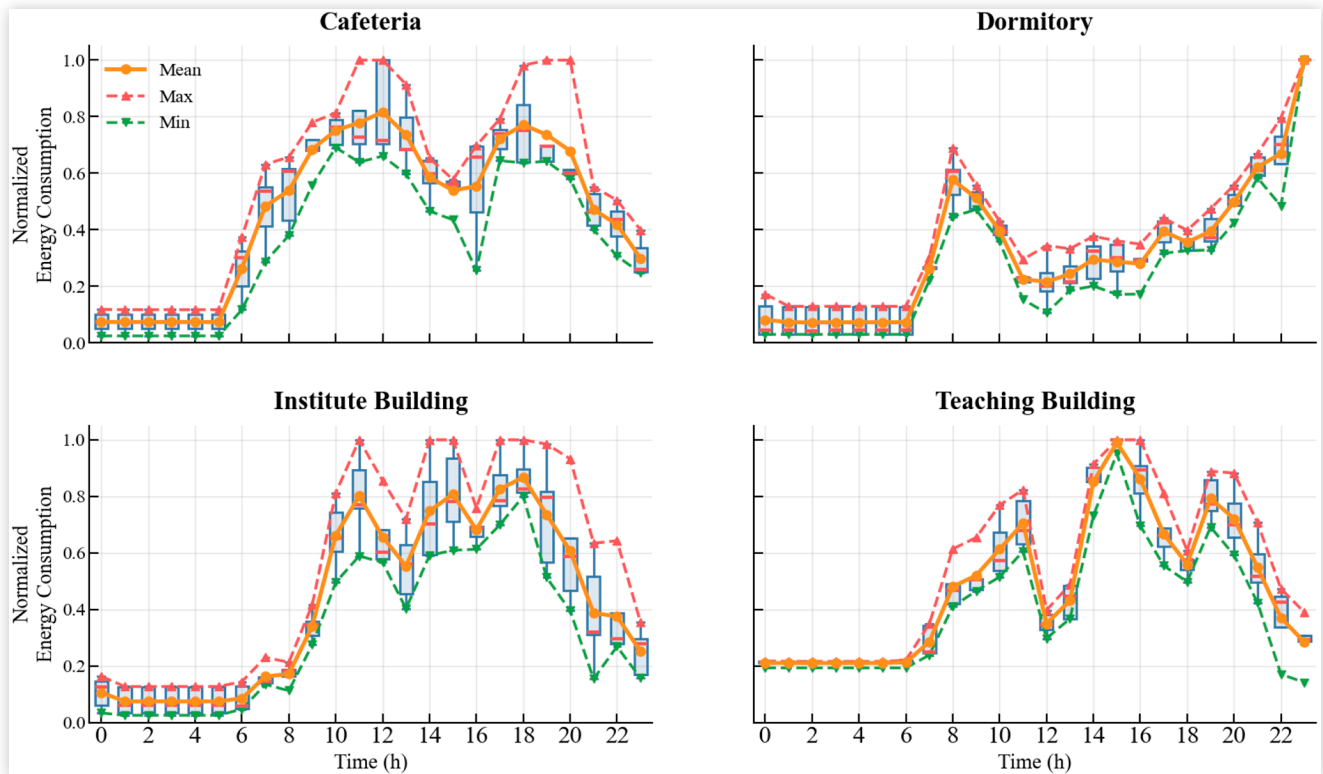
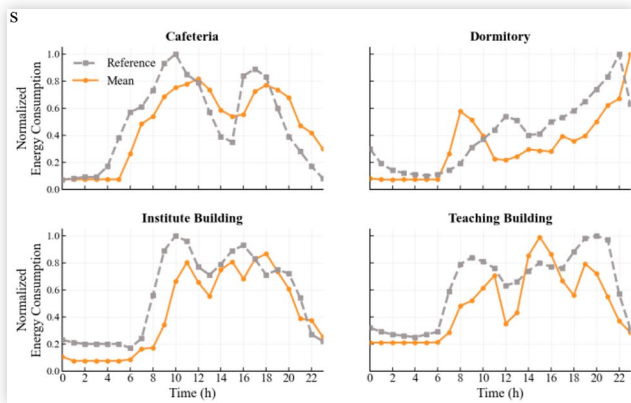


FIGURE 7 Comparison of simulated values and actual energy consumption data for the four campus buildings (Cafeteria, Institute Building, Teaching Building, and Dormitory)



that a complete 24-hour simulation in the experiment required the consumption of 14,193,144 tokens (including both prompt and response tokens). Given the limitations of computational resources, significantly higher memory resources and time costs would be incurred if 495 LLM Agents were simulated via a fully decentralized approach. It is projected that, if the same simulation duration is achieved under such a full-agent scheme, over 200 million tokens would be required, and the experiment time would increase exponentially. In conclusion, computational resource consumption is reduced to approximately 6.26%

by the proposed "Archetype-Individual" scheme, while high-fidelity results approximating real-world outcomes are successfully maintained.

Conclusions

This study proposes the Hierarchical Generative Agent-based Simulation Framework (HGA-Sim) to address the inherent complexities and limitations of traditional modeling approaches in capturing the nuanced, heterogeneous, and adaptive nature of human behavior in sustainable mobility and energy systems. The urgent need for such a novel research paradigm arises from the significant uncertainties that system design and policy evaluation often encounter due to the simplifying assumptions about human decision-making prevalent in existing models.

The HGA-Sim framework overcomes traditional bottlenecks through two major core innovations. Firstly, it leverages the LLM to autonomously generate agents endowed with distinct personality traits, capable of simulating diverse human-like responses to environmental influences and experiences. This approach effectively overcomes the limitations of stochastic simplifications in conventional models, enabling the automatic generation of personalized user profiles and significantly reducing the laborious and subjective manual configuration previously required for large-scale simulations. Furthermore, implementing generative agent systems at community

TABLE 4 Comparison of HGA-Sim with Traditional Methods

Dimension	Traditional Rule-Based ABM	Proposed HGA-Sim (Generative ABM)	Advantages of HGA-Sim
1. Decision-Making Logic	Predefined Rules & Heuristics: Agents follow hard-coded "If-Then" logic or probabilistic schedules (e.g., Markov Chains).	Autonomous Reasoning: Agents use LLMs to "think" and generate actions based on perception, memory, and reflection.	Captures nuanced, non-linear human decision-making processes that rules cannot cover.
2. Configuration Complexity	High/Manual: Requires labor-intensive manual setting of rules and schedules for each agent type.	Low / Automated: "User Personas" are automatically generated via LLMs based on psychological traits (Big Five).	Significantly reduces setup time and subjectivity for large-scale simulations.
3. Behavioral Heterogeneity	Limited: Often uses homogenous templates for groups; difficult to simulate unique individual quirks.	High: Each agent acts differently based on unique personality traits and memory streams.	Reproduces realistic diversity and randomness in community behaviors.
4. Adaptability (New Scenarios)	Low: Fails or requires re-coding when facing undefined scenarios (e.g., unexpected outages).	High (Zero-Shot): Agents can use common-sense reasoning to adapt to novel environmental stimuli without re-training.	Better suited for testing future policies or emergency responses.
5. Data Dependency	High: relies heavily on large-scale historical datasets to train probability distributions.	Moderate: Can function with high-level descriptions (Prompts) and minimal seed data.	Viable in data-scarce environments where privacy concerns limit data collection.

or even city scales is currently almost impossible without further breakthroughs in LLMs' inference and computational demands [11]. Therefore, the framework adopts an innovative "Archetype-Individual" hierarchical architecture. By clustering individual agents into representative archetypes, this architecture dramatically enhances the computational feasibility of large-scale community-level simulations. While significantly reducing the invocation costs of large language models, this architecture ingeniously preserves the essential heterogeneity of individual behaviors through probabilistic sampling techniques, achieving an optimal balance among simulation scale, economic viability, and the fidelity of mobility and energy consumption patterns.

The efficacy and practical value of the HGA-Sim framework were rigorously validated through an empirical case study of a campus community comprising 495 agents. The simulation accurately replicated macroscopic mobility and energy consumption patterns, including critical peak loads and temporal dynamics, exhibiting remarkable agreement with real-world data (RMSE = 0.1983). This success highlights the framework's ability to serve as a foundational tool for assessing the feasibility of sustainable mobility solutions, quantifying the socio-economic impacts of energy policies, and mitigating investment risks in sustainable infrastructure. It achieves this by creating a human-in-the-loop virtual testing environment.

For future work, the HGA-Sim framework can be further enhanced by incorporating more complex, realistic scenario designs and moving beyond single-day simulations to explore extended periods. This includes analyzing the distinct behavioral patterns and energy consumption profiles between weekdays and weekends, as well as investigating seasonal variations in human activity and resource usage. Additionally, future research could explore the integration of more dynamic external factors and refine the mechanisms of interaction between

archetypes and individuals to capture even more nuanced behavioral patterns.

References

1. IEA, "International Energy Agency," IEA n.d. <https://www.iea.org>. Accessed 27 Sept. 2025.
2. Socolow, R., "Twin Rivers Program on Energy Conservation in Housing: Highlights and Conclusions," *Office of Scientific and Technical Information (OSTI)* 1 (1978), doi:10.2172/6607206.
3. Evins, R., "A Review of Computational Optimisation Methods Applied to Sustainable Building Design," *Renewable and Sustainable Energy Reviews* 22 (2013): 230-245, doi:10.1016/j.rser.2013.02.004.
4. Ding, Y. et al., "An Occupancy-Based Model for Building Electricity Consumption Prediction: A Case Study of Three Campus Buildings in Tianjin," *Energy and Buildings* 202 (2019): 109412, doi:10.1016/j.enbuild.2019.109412.
5. Jin, Y. et al., "Building Occupancy Forecasting: A Systematical and Critical Review," *Energy and Buildings* 251 (2021): 111345, doi:10.1016/j.enbuild.2021.111345.
6. Gao, C. et al., "Large Language Models Empowered Agent-Based Modeling and Simulation: A Survey and Perspectives," *Humanities and Social Sciences Communications* 11, no. 1 (2024), doi:10.1057/s41599-024-03611-3.
7. Ghaffarzadegan, N. et al., "Generative Agent-based Modeling: An Introduction and Tutorial," *System Dynamics Review* 40, no. 1 (2024), doi:10.1002/sdr.1761.
8. Park, J.S. et al., "Generative Agents: Interactive Simulacra of Human Behavior," *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (2023): 1-22, doi:10.1145/3586183.3606763.

9. Williams, R. et al., "Epidemic Modeling with Generative Agents," arXiv preprint arXiv:2307.04986, 2023, doi:10.48550/arXiv.2307.04986.
10. Argyle, L.P. et al., "Out of One, Many: Using Language Models to Simulate Human Samples," *Political Analysis* 31, no. 3 (2023): 337-351, doi:10.1017/pan.2023.2.
11. Li, N. et al., "EconAgent: Large Language Model-Empowered Agents for Simulating Macroeconomic Activities," *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (2024): 15523-15536, doi:10.18653/v1/2024.acl-long.829.
12. Almashor, M. et al., "Can Private LLM Agents Synthesize Household Energy Consumption Data?," in *The 15th ACM International Conference on Future and Sustainable Energy Systems*, 2024, 664-68, doi:10.1145/3632775.3661993.
13. "Cloud Models are Now Available in Ollama Chat & Build with Open Models," *Ollama*, www.ollama.com. Accessed 29 Sep. 2025.
14. "Qwen2.5-7B," *Hugging Face*, <https://huggingface.co/Qwen/Qwen2.5-7B>. Accessed 29 Sep. 2025.
15. Jianan, Li, "The Analysis and Strategy of Public Building Energy Consumption in Shandong Jianzhu University," Shandong Jianzhu University, MA thesis. 2015, doi:10.27273/d.cnki.gsajc.2024.001009.

Contact Information

Shiqi(Shawn) Ou

sou@scut.edu.cn; oushiqi@pazhoulab.cn

Phone number: +86-020-81181684

Mailing address:

1. South China University of Technology, School of Future Technology, 777 Xingye Ave East, Panyu District, Guangzhou, Guangdong, 511442, China

2. Guangdong Artificial Intelligence and Digital Economy Laboratory (Guangzhou), 70 Yuean Road, Haizhou District, Guangzhou, Guangdong 510335, China.

Acknowledgments

This research is supported by the Introduced Innovative R&D Team of Guangdong (2023ZT10L145). All opinions expressed in this paper are the authors' and do not necessarily reflect the policies and views of sponsors.

Appendix

Examples of Prompts and Outputs for Agents' Planning Module

Here, the operational logic of the Planning Module is illustrated through two specific examples. First, a daily coarse-grained plan for an agent is generated. Second, the decomposition of this daily summary into specific hourly tasks is presented. In this process, a plan for a specific hour is broken down into minute-level activities, each with a distinct duration. When the agent's status involves an interaction with energy-consuming devices, this duration is parsed and used in the subsequent calculation of energy consumption.

Prompt:

``${base_desc}``

Here is the typical hourly breakdown for ``${archetype}`` today:

``${daily_schedule}``

Please refer to the above archetype information and daily schedule to generate an hourly plan (24-hour format). Only fill in the `<activity>` content, do not skip any time point.

The hourly plan must follow the format below:

````

``${hourly_schedule}``

````

Output:

schedule:

00:00~01:00: <sleeping>:

01:00~02:00: <sleeping>:

02:00~03:00: <sleeping>:

03:00~04:00: <sleeping>:

04:00~05:00: <sleeping>:

05:00~06:00: <sleeping>:

06:00~07:00: <sleeping>:

07:00~08:00: <sleeping>:

08:00~09:00: <breakfast>:

09:00~10:00: <attending campus events or group projects>:

10:00~11:00: <attending campus events or group projects>:

11:00~12:00: <attending campus events or group projects>:

12:00~13:00: <lunch>:

13:00~14:00: <working on research proposal>:

14:00~15:00: <working on research proposal>

15:00~16:00: <preparing for seminar presentation>

16:00~17:00: <preparing for seminar presentation>

17:00~18:00: <rainy day, stay in apartment and read related articles>

18:00~19:00: <rainy day, stay in apartment and read related articles>

19:00~20:00: <watching TV or relaxing>

20:00~21:00: <watching TV or relaxing>

21:00~22:00: <watching TV or relaxing>

22:00~23:00: <watching TV or relaxing>

23:00~00:00: <going to bed>

Prompt:

`\${base_desc}`

Decompose the plan into sub-tasks.

The current typical plan for `\${archetype}` is: `\${plan}`

The human sensation of the weather: `\${weather}`

The total number of sub-tasks should not exceed 10, with each sub-task on a separate line in the following

Output format:

1) `\${archetype}` *plans* to* <activity> (Duration: <time>, Remaining: <time>)

for example:

1) Archetype1 *plans* to review kindergarten curriculum standards (duration: 15, remaining: 45)

2) Archetype1 *plans* to brainstorm for the lesson (duration: 10, remaining: 35)

3) Archetype1 *plans* to create the lesson plan (duration: 20, remaining: 15)

4) Archetype1 *plans* to print the lesson plan (duration: 10, remaining: 5)

5) Archetype1 *plans* to put the lesson plan in her bag (duration: 5, remaining: 0)

The duration of each sub-task must be an integer multiple of `\${stride}` minutes, think step by step, list all subtasks for `\${archetype}` between `\${start}` and `\${end}` (total duration: 60 minutes!):

<subtasks>

Output:

schedule:

00:00-01:00: <sleeping>

01:00-02:00: <sleeping>

02:00-03:00: <sleeping>

03:00-04:00: <sleeping>

04:00-05:00: <sleeping>

05:00-06:00: <sleeping>

06:00-07:00: <sleeping>

07:00-08:00: <sleeping>

08:00-09:00: <breakfast>

08:00-08:20: to have breakfast

08:20-08:35: to review the research proposal outline

08:35-08:45: to gather necessary research materials and references

08:45-08:55: to draft the introduction section of the research proposal

08:55-09:00: to prepare for the upcoming seminar presentation on renewable energy technologies

09:00-10:00: <attending campus events or group projects>

09:00-09:25: to attend campus events or group projects

09:25-09:35: to review research proposal notes

09:35-09:50: to outline the seminar presentation on renewable energy technologies

09:50-10:00: to prepare key points for the seminar presentation

10:00-11:00: <attending campus events or group projects>

10:00-10:20: to finalize the research proposal outline

10:20-10:35: to review literature sources for the proposal

10:35-10:45: to draft the introduction section of the proposal

10:45-10:55: to prepare an outline for the methodology section

10:55-11:05: to review and organize notes for the upcoming seminar presentation

11:00-12:00: <attending campus events or group projects>

11:00-11:15: to finalize the research proposal outline

11:15-11:25: to review literature sources for the proposal

11:25-11:35: to draft the introduction section of the proposal

11:35-11:45: to research and compile relevant data for the proposal

11:45-11:55: to draft the methodology section of the proposal

11:55-12:00: to review and revise the drafted sections

12:00-13:00: <lunch>

12:00-12:15: to work on research proposal - introduction section

12:15-12:35: to work on research proposal - literature review

12:35-12:45: to take a short break and grab some water

12:45-12:55: to work on research proposal - methodology section

12:55-13:00: to review the progress made so far and make notes for next steps

13:00-14:00: <working on research proposal>

13:00-13:25: to work on research proposal

13:25-13:30: to take a short break and review the weather

13:30-13:50: to continue working on research proposal

13:50-14:00: to finalize sections of the research proposal

14:00-15:00: <working on research proposal>

15:00-16:00: <preparing for seminar presentation>

16:00-17:00: <preparing for seminar presentation>

17:00-18:00: <rainy day, stay in apartment and read related articles>

18:00-19:00: <rainy day, stay in apartment and read related articles>

19:00-20:00: <watching TV or relaxing>

20:00-21:00: <watching TV or relaxing>

21:00-22:00: <watching TV or relaxing>

22:00-23:00: <watching TV or relaxing>

23:00-00:00: <going to bed>

Examples of Prompts and Outputs for Agent Behavioral Decision-Making

This section illustrates the prompt strategy employed for agent decision-making. By embedding the scheduled "activity" from the plan into the prompt context, the agent can make decisions about interactions with devices currently perceived. The output shown in the example, "computer148," represents the specific computer device designated as number 148 within the designed campus environment.

Prompt:

Select the most relevant object for the current activity from the options list.

Current activity: $\{activity\}$

Requirements:

1. Must select an object from this list: $\{objects\}$.
2. Do not select objects outside the list.
3. Output the selected result directly.

The most relevant object for the current activity is:

Output:

computer148