Acta Polytechnica CTU Proceedings **57**:263–274, 2026
https://doi.org/10.14311/APP.2026.57.0263

Published by the Czech Technical University in Prague

GENERATIVE AGENTS IN CROWD SIMULATION: A COGNITIVE APPROACH WITH LARGE LANGUAGE MODELS

Nizar Ntarouis*, Roland Geraerts

Utrecht University, Heidelberglaan 8, 3584 CS Utrecht, The Netherlands

* corresponding author: nizardarwish001@gmail.com

Abstract. Crowd simulation is a powerful tool used in fields such as urban planning, emergency response, and entertainment to model and predict human movement and behavior in various scenarios. As society becomes increasingly complex and interconnected, the need for simulations that accurately capture human behavior at both the individual and group level grows. Understanding these interactions can help institutions and experts develop more effective mitigation strategies in dynamic social environments.

This research explores the potential of the power of Large Language Models (LLMs) in crowd simulations, leveraging their capabilities to model individual behavior and enable the emergence of realistic crowd dynamics through agent-level interactions.

We propose a novel architecture that integrates key cognitive components – such as perception, planning, memory, reflection, and action – on an algorithmic level. This approach allows generative agents to process environmental and social contexts in a human-like manner. Our findings show that these agents exhibit diverse and contextually appropriate behaviors, closely resembling human decision-making, particularly in crisis situations.

Keywords: Crowd behavior, Large Language Models, cognitive models, crisis scenarios.

1. Introduction

In the rapidly advancing field of crowd simulation, emulating the complexity of human interaction is essential, as its fidelity has profound implications for crisis management, urban planning and public safety. Central to this quest is the process of pattern extraction through the concept of Patterns of Life (PoL) [1], which refers to the repetitive and structured aspects of existence, whether societal, individual, or organizational level. These patterns may involve cycles, rhythms, or structures that are recurrent and contribute to the overall order and functioning of a particular system.

Large Language Models (LLMs) offer a transformative capability in this regard, enabling the nuanced simulation of individual behaviors and interactions [2, 3]. Our goal is to study patterns related to observation and analysis of behaviors in crisis and evacuation scenarios. This includes examining the thought processes of individual actors, as well as the interactions and collisions that occur both internally (within the actors' minds and their mental models of the world) and externally with nearby actors.

Traditional Methodologies and PoL research often fall short, oversimplifying human cognition and complexities by generalizing and focusing on clusters of group movement and navigation [4]. These methods typically overlook subtleties of individual cognition -the thoughts, decisions, and interactions that result in the passive formation of crowds. Understanding crowd dynamics involves recognizing emergent properties that arise from individual interactions but are not inherent in isolated individuals (see Figure 1).

In contrast to conventional approaches that prescribe behavioral rules at the group level, treating agents as largely interchangeable, our approach models crowd behavior from the cognitive perspective of each individual. Agents reason through a unique profile comprising their mental state, emotions, personality, interests, personal perspectives, and social associations. This means that rather than encoding behavior as fixed rules, the LLM reasons about these attributes in context, producing decisions that reflect each agent's individual circumstances. For instance, a frightened agent may still choose to re-enter a building to search for a friend, while another with the same level of fear but no social ties might evacuate immediately. This cognitive grounding allows high-level behaviors and crowd-level patterns to emerge naturally from individual decision-making, rather than being imposed from the top down.

In our exploration of the vast network of human behavior, we pose the question: Can Large Language Models be used to simulate realistic crisis scenarios? This inquiry unfolds into two avenues of exploration, which underpin aspects and challenges of the complex network of interplay that defines crowd behavior:

- Can LLM-based agents exhibit human-like reasoning and communication, leading to emergent behavioral patterns that closely resemble those observed in real-world human interactions?
- What are the performance and scalability limita-

Figure 1. Festival: Generative agents mingle, form clusters, and navigate through a vibrant festival, showcasing the emergent complexity of crowd dynamics.

tions of LLM-based simulations, particularly in complex scenarios involving large populations?

2. Methodology

Achieving the creation of artificial agents capable of natural language interactions and emergent social behaviors [5, 6] represents a significant challenge in artificial intelligence, particularly at the micro level of human behavior. This study leverages recent advancements in generative AI and cognitive models, centering around a Large Language Model (LLM) as the primary architecture controller [7]. The architecture integrates cognitive modules that enable agents to navigate environments and facilitate communication between the LLM and geometric simulations. We aim to determine whether generative agents within this setup could effectively coordinate, form relationships, carry out activities, and respond dynamically in a virtual context.

We propose a novel architecture for crowd simulation that integrates cognitive components, such as perception, planning, memory, reflection, and action, on an algorithmic level to enable generative agents to process environmental and social contexts in a human-like manner [8]. Inspired by dual-process theories of cognition [9], we deploy two systems embedded inside the agentic framework: the Intuitive and Deliberative systems.

2.1. Two thinking systems

System 1 operates automatically, driven by heuristics from previous experiences and emotions. In the context of generative agents, it is activated for events requiring immediate responses, using predefined symbolic decisions to each agent's personality. In other words, it is both an objective and subjective experience, where events like an earthquake or a shooting scenario would be considered of high importance by almost all agents. In contrast, an announcement about a lecture being canceled will only concern a subset of student agents.

Upon activation, System 1 leverages predefined symbolic decisions to react instinctively. In emergency scenarios, such as a fire breakout, the framework would immediately trigger a flight response without any planning or analysis of the situation.

System 2 involves deliberate and logical processing, requiring more cognitive resources. Representing the LLM architecture, System 2 is responsible for analyzing significant events and constructing rational responses. This system enables agents to process information, consult memories, and develop plans effectively.

Generative agents operate within dynamic environments, influenced by current settings and past experiences. This study proposes an agent architecture that integrates a large language model with mechanisms for gathering, processing, and utilizing information.

2.2. Architecture design

Generative agents are designed to function within dynamic environments, engaging with other agents and adapting to environmental changes. The architecture merges large language models with mechanisms that enable the utilization of their surroundings and memories to generate actions.

To bridge the gap between LLMs and autonomous agents, it is necessary to design rational auxiliary components to assist LLMs to perform in a human-like manner. In this direction, previous work has developed a number of modules to enhance LLMs [6, 10–14]. These components, inspired by human cognition, are integrated on a computational and algorithmic level [15].

In complex agent-based simulations, another challenge is to ensure that all processes and components, although interconnected, can operate independently and efficiently. Thus, we propose an asynchronous multi-threaded framework that allows different modules to work in parallel, improving responsiveness and scalability. Asynchronous processing allows the system to handle multiple agents simultaneously without being bottlenecked by any single module, improving overall performance while generating contextually grounded outputs in real-time.

In this section, we analyze this asynchronous architecture (Figure 2), comprising a profiling module,

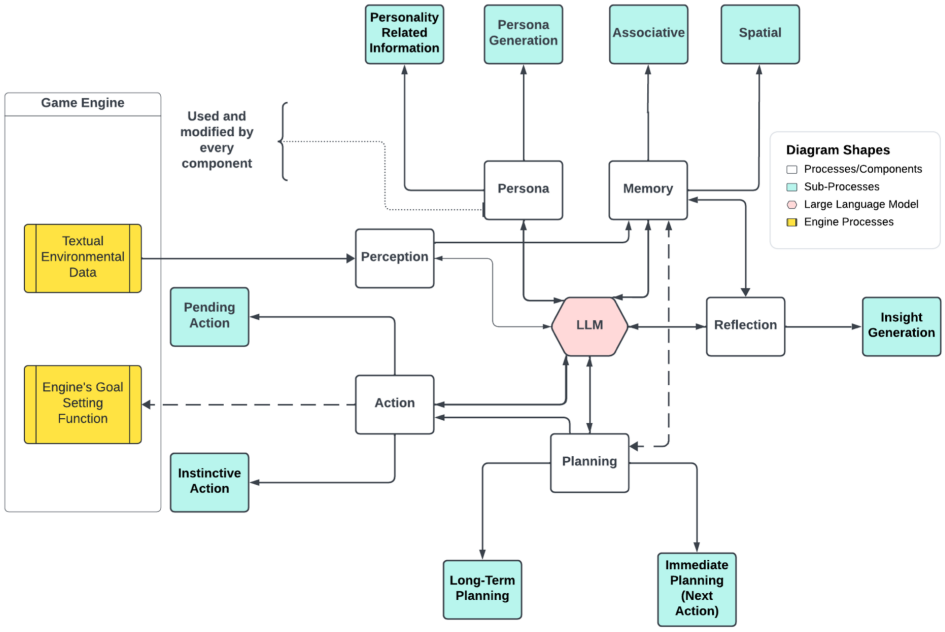

FIGURE 2. The Component Diagram shows the interconnections between architecture components. Solid arrows indicate direct connections where one component's execution always effects the other. Dashed arrows indicate conditional connections where influence may occur but not in every instance.

memory module, reflection module, planning module, and action module, all working in harmony to enable for real-time dynamic simulations, critical for multi-agent environments.

This framework allows generative agents to exhibit sophisticated behaviors that closely resemble human cognition, enhancing the realism and applicability of simulations. As a result, agents naturally form group patterns that reflect real-world human behavior.

2.3. PROFILING MODULE

Agent personas are generated through a hierarchical sampling process combining automatic generation via prompts and optional dataset alignment. The user specifies the scenario context, desired population size, and subgroups (e.g., Staff, Attendees). The population is then split into subgroups with LLM-generated percentage distributions (e.g., 10 % Staff, 85 % Attendees, 5 % Bands). Each subgroup is further divided into social groups with interpersonal relationships (e.g., 30 % Couples, 20 % Families, 10 % Solo Attendees). Personas are generated in random batches of 2 to 5 per social group, allowing the LLM to establish relationships within each batch. The resulting profiles, stored in JSON format, can be reviewed and modified by the user. Optionally, few-shot examples from real-world datasets guide the generation towards more realistic personas. The population generation and persona creation prompts are provided in the supplementary materials [16].

2.4. PERCEPTION MODULE

In generative agent architectures where language models serve as central controllers, translating environmental state into textual representations presents a fundamental challenge. The perception module bridges this gap, enabling agents to process and interpret their surroundings through two primary channels: textual event streams and spatial entity information. Operating at the simulation loop frequency (i.e., 10 Hz), the module provides real-time environmental awareness, maintaining computational efficiency through selectively processing novel information alone. In consequence, components activate conditionally rather than synchronously (Figure 3).

The simulation environment is represented through a hybrid symbolic spatial encoding. Events are textual descriptions of actions, state changes, or environmental conditions (e.g., "bar is closed", "fire alarm triggered", "agent_0 is running"). Objects are labeled by the user with associated textual descriptors and event annotations that can be injected dynamically during runtime. Agents are labeled with unique identifiers (agent_0, agent_1, etc.), and their actions are described textually by the system. Interactions fall into three categories: agent-object interactions ("agent_0 is using laptop"), agent-agent interactions ("agent_0 is chatting with agent_1"), and generic actions ("agent_0 is walking").

Spatial awareness is implemented through radius-based perception: agents observe all entities within a configurable distance threshold, receiving both their absolute positions and relative distances to establish 2D depth perception. This textual encoding circumvents the need for visual processing, though emerging multimodal architectures with video processing capabilities [17] represent a promising direction for pixel-level environmental understanding.

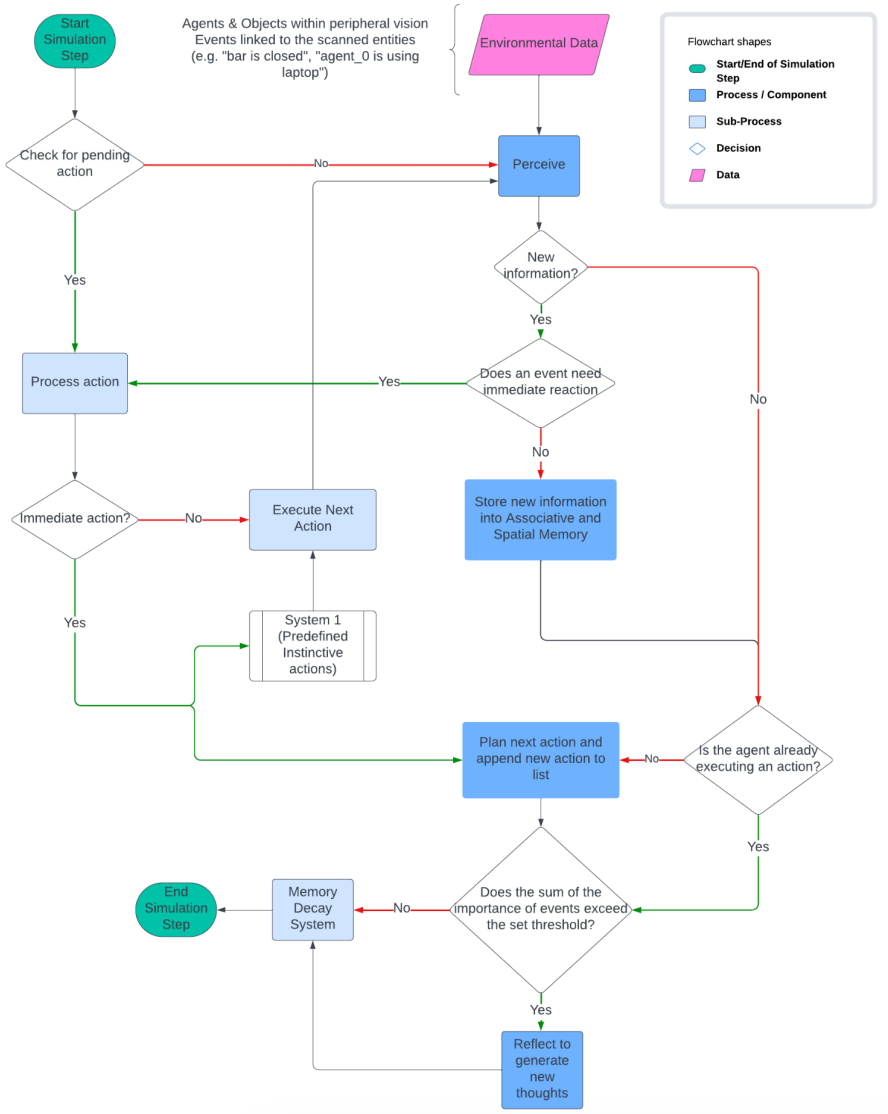

Figure 3. Agent decision making flowchart.

The module executes a multi-stage processing pipeline at each simulation tick:

(1.) **Information Gathering and Filtering:** All entities within the perceptual radius are compared against recently observed information in memory. New observations are processed and stored, whereas existing entities have their recency scores updated, minimizing redundant processing.

(2.) **Event Classification and Importance Scoring:** Perceived events are classified as dangerous or attractive and assigned importance scores from 1 to 10 (see Event Classification prompt in supplementary materials [16]). These scores depend on persona-specific attributes, environmental, and inner state. For instance, in a concert scenario, agents would deem "stage pyrotechnics malfunction" as objectively dangerous (high importance across all agents), while "indie band starts playing" would receive subjective ratings: high importance for fans of that genre, low importance for others who may prioritize finding refreshments or socializing.

(3.) **Threshold-Based System Activation:** Importance scores determine which cognitive system activates:

- **High Importance ($\geq$8):** Triggers System 1, interrupting ongoing activities to focus on critical events. System 2 is activated concurrently and overwrites System 1's action once the LLM completes its computation.
- **Moderate Importance ($\geq$4):** Updates emotional and physical status without interrupting current actions. For example, hearing preferred music at a festival may shift an agent's state from "bored" to "excited."

(4.) **Memory Encoding:** Events are converted to embedding vectors that capture their semantic meaning. These embeddings, along with textual descriptions, are stored in associative memory. Previous events related to the same entity are updated to past tense, maintaining temporal consistency.

The perception module plays a critical role in protecting the architecture from vulnerabilities common to language models, namely prompt injection, misinformation, and adversarial attacks [18]. Employing selective attention, filtering by spatial proximity, importance thresholds, and temporal novelty balances the complexity of managing computational costs and context window limitations while preserving behavioral realism. This approach ensures that information reaching memory or planning components stays relevant and accurate, yet maintaining awareness.

2.5. Memory module

Generative agents that emulate human behavior must reason about past and current experiences to inform future planning. However, the volume of memories generated from perceived events far exceeds the limitations of typical LLM context windows. Moreover, not all stored information remains relevant to an agent's current state. Thus, efficient memory retrieval mechanisms are essential in selecting contextually relevant memories that enable the most informed decision-making at any given moment.

The memory module maintains a comprehensive record of each agent's experience, encompassing perceived environmental information, inter-agent interactions, and insights derived from reflection on past observations. As illustrated in Figure 2, the perception, reflection, and planning components, both utilize and generate memories. This bidirectional relationship enables agents to self-evolve, exhibiting increasingly consistent, reasonable, and subjective behavior.

2.5.1. Memory architecture

Information provided by the perception module is stored in dual representations: textual and vector-encoded formats. We partition memory into two distinct streams:

Spatial Memory: Stores perceived entities (objects and agents) within the agent's current visual radius, along with their respective positions. The visual radius is a parameter that the experimenter can adjust.

Associative Memory: Maintains events and relational information associated with perceived entities.

Selective and efficient retrieval is critical for real-time decision-making based on environmental state, agent goals, past experiences, and contextual factors. We implement two complementary retrieval methods:

Keyword Matching: Applied to object-related memories, providing direct but less nuanced access. By matching currently perceived entities and stored linguistic information retrieves memories explicitly linked to objects or agents.

Vector Space Retrieval: Employs a scoring function comprising three weighted components: recency (assigns higher scores to recently accessed memories), importance (context-dependent significance assigned by the perception module), and relevance (computed via cosine similarity between stored memory embeddings and query embeddings).

The retrieval score normalizes all components to $[0, 1]$ via min-max scaling, computing a weighted combination (see Algorithm 1 in supplementary materials [16]):

$$\text{score} = \alpha_r \cdot r + \alpha_v \cdot v + \alpha_i \cdot i, \tag{1}$$

r, v, and i represent recency, relevance, and importance. In this study, all weights are set equally ($\alpha_r = \alpha_v = \alpha_i = 1$), as the focus on short-duration crisis scenarios, combined with the forgetfulness function that actively prunes older memories, results in a small and already relevant memory pool where differential weighting has limited impact. Top-ranked memories are subsequently incorporated into the LLM prompt.

Forgetfulness Function: Agent-based modeling frameworks commonly encounter memory limitations when scaling to large populations [19], as the information space grows exponentially with entity count. To address this, we implement a time-based forgetfulness function. Given our focus on crisis scenarios, inherently short-duration events, we establish a one-minute temporal threshold: memories unaccessed by any component within this window are discarded. This hyperparameter remains user-adjustable for alternative simulation contexts.

The memory module, inspired by human cognitive processes and implemented at the algorithmic level [15], proves essential for realistic agent behavior. By selectively retrieving and pruning information based on recency, relevance, and importance, our approach enables agents to adapt dynamically to evolving contexts. The resulting memory system enhances both the realism and computational efficiency of our simulation framework, particularly in resource-constrained scenarios involving large agent populations.

2.6. Planning module

Effective planning requires navigating immediate circumstances and long-term objectives. Agents must generate plans, act on them through intermediate steps, and revise based on environmental feedback and interactions.

The planning module aims to empower agents with three human-like capabilities: long-term planning, task decomposition, and immediate planning:

Long-Term Planning: Every 24 hours, agents plan their day and create a schedule. Each agent's plan varies significantly based on daily requirements, goals, lifestyle, physical and emotional status, and external factors such as weather conditions or environmental obstacles. Agents construct 24-hour plans with activities lasting at least one hour.

Task Decomposition: Each activity is decomposed into plausible sub-tasks in 5-minute increments. For example, "work for 1 hour" might break down into

"respond to emails for 15 minutes", "attend team meeting for 30 minutes", and "review project documents for 15 minutes."

Immediate Planning: This represents the most critical split-second decision an agent makes, determining the next immediate action. The agent considers current time, internal state, persona, relevant memories, recent environmental observations, the general plan, previous actions, and the current task. The agent can either continue with the current task or diverge from it. For instance, at a music festival, the immediate plan could be grabbing food to address hunger, going to the bathroom to relieve discomfort, or making split-second decisions in response to emergent dangers like nearby fire or an earthquake. This flexibility allows agents to adapt plans dynamically based on new information.

The planning process retrieves relevant memories, decides on the next immediate plan, and stores it in memory to be reflected upon. This plan is provided to the action module to determine the course of action and new location. The daily planning and task decomposition prompts are provided in the supplementary materials [16].

Our architecture adopts a middle ground between planning with and without environmental feedback. Agents do not directly alter planning based on external reactions, but generate insights through the reflection component that may influence future plans. Feedback affects agent variables such as emotional state, physical condition, position, and reflections, which can disrupt or reinforce previous plans during subsequent planning cycles, fine-tuning behavior to react appropriately in a believable, human-like manner. To fundamentally change an agent's thinking would require altering the memories and beliefs they intrinsically hold.

2.7. Action module

The action module translates agent decisions into specific outcomes at the most downstream level of the architecture. Directly influenced by the planning module, it connects to the engine's goal-setting function to interact with the environment (Figure 2). It is important to note that this work focuses on the cognitive and behavioral aspects of crowd simulation, not on path planning. The LLM-based architecture produces high-level decisions, such as goal positions and intended actions, which are then passed to a crowd simulation framework [20] that handles path planning, trajectory computation, and collision avoidance. Example prompts for room selection, object choice, and immediate action decisions are provided in the supplementary materials [16]. This section examines the action module from four perspectives: action goal, action production, action space, and action impact.

Action Goal: Agents perform three primary functions. First through task decomposition, actions break down into tangible steps, allowing agents to determine which objects fulfill their needs and goals. Second, they communicate with each other, discuss topics, form relationships, and coordinate through conversation. Third, they navigate the environment, discovering new areas and objects, expanding their memory streams and perceptual understanding.

Action Production: In our architecture, action production draws from two strategies: action via memory recollection and action via plan following [10, 21, 22]. While considering the decomposed plan, immediate plan, previous actions, and in-context memories, it retains flexibility to diverge from instructions. This mirrors human behavior, where thoughts do not always directly translate into actions.

Action Space: When choosing actions, agents decide whether to move to an area or entity through planning [6, 23], chat with other agents [24], remain idle, or use an object.

Action Impact: Our implementation incorporates two impact types: internal state alteration and action triggering. When an action is taken, memory streams are updated and physical, emotional, and persona attributes adjust in response to outcomes. For instance, if an agent engages in positive conversation, it updates its emotional state to reflect increased happiness and stores this positive memory.

Critically, actions also impact surrounding agents who perceive these behaviors. When nearby agents observe one another such as, an agent fleeing from danger or engaging in conversation, they store this observation in their own memory streams, potentially triggering their own planning and action responses. This creates emergent crowd dynamics where individual behaviors propagate through the population.

This ripple effect, where each reaction establishes the framework for future contextual behavior, passively creates a dynamic world resembling real-life scenarios. For example, if an agent experiences fear after a dangerous situation, it adapts by recalling this memory in similar circumstances, enabling better informed decisions in the future. Simultaneously, agents observing this fearful response may infer danger and adjust their own behavior accordingly, leading to collective patterns like crowd evacuation without explicit coordination.

2.8. Reflection module

Perceived data in multi-agent simulations grows exponentially with the number of entities and objects in the environment. Although raw observational data enables informed decisions, excessive information can obscure what is actually important or relevant, making it difficult for the LLM to extract meaningful stored information during planning. Agents struggle to generalize, draw conclusions, and form higher-level inferences from streams of low-level observations.

To address this, we implement a reflection module that generates higher-level insights, inferences, and associations between pieces of information. These

reflections provide abstract representations of raw data and are generated periodically whenever the cumulative importance of new observations reaches a set threshold. The insights are stored in associative memory and retrieved by other components (Figure 2), or become the subject of further, even more abstract reflections.

For instance, an agent observing various agents in a particular context might reflect: "This place seems popular and worth checking out". The decision then triggers future reflections making new inferences that reinforce or contradict previous conclusions. This reflective process enhances the agent's subjective experience and ability to make adaptive choices in dynamic environments.

The reflection process (see Algorithm 2 in supplementary materials [16]) generates high-level questions based on recent experiences. For example, if records include "fire in the concert hall" and "agent_11 is near the fire", potential questions might be: "What emergency situations are occurring and how are agents responding?" or "What actions are being taken to ensure attendee safety and security?". These questions retrieve relevant memories, and the LLM extracts insights such as "The fire is expanding uncontrollably" or "Several agents are attempting to contain the fire" which directly influence immediate decision-making. Finally, insights are classified by importance, converted to vector representations, and stored in memory.

Over time, agents construct cognitive maps where high-level thoughts are used to generate increasingly abstract representations. Even when raw observations are eventually forgotten, the insights derived from them persist, shaping long-term behavior. The associated reflection prompts are provided in the supplementary materials [16].

3. Evaluation

To determine whether agents exhibit human-like decision-making across multiple scenarios, we employ four evaluation methods:

1. **Pattern Comparison (Subjective Evaluation):** We compare behavioral patterns on micro and macro levels. By aligning data from simulated scenarios with expert assessments of real-world evacuation patterns [25–30], we evaluate the realism and effectiveness of agent actions from both individual and crowd perspectives.

2. **Crowd Scalability:** We test the architecture's boundaries by examining the correlation between agent population and real-time model responsiveness.

3. **Agent Interview:** We leverage generative agents' ability to communicate in natural language by conducting interviews. We analyze their their perceptual awareness, ability to plan, recall, reflect, reason about past experiences and respond to events in their simulated lifetime.

4. **Patterns of Life Scoring:** Utilizing Silverman's scoring matrix [31], we determine the category of our simulation based on three key dimensions: activities, relations, and cognition.

This examination validates whether behaviors emerging from the model resemble those of humans. In the following sections, we delve into the emergent human-like behaviors observed, along with the specifics of the evaluation strategies.

3.1. Subjective evaluation

We conducted a within-subjects study with 25 participants. Participants watched replays of agent actions in controlled environments and answered questions based on their observations. These environments examine specific behavioral patterns in narrowly defined contexts, testing whether agents exhibit social influence, familiarity with exits, exit signage influence, exit allocation, emotional reactions, and coordination.

Participants provided feedback on four aspects: (1) pattern comparison accuracy during evacuation and crisis situations, (2) realism of agent behaviors, (3) emotional intelligence and its influence on decision-making, and (4) diversity of behaviors compared to human populations.

3.1.1. Experimental process

A comprehensive questionnaire (see supplementary materials [16]) was designed to guide participants through the evaluation. Each section provided scenario context, video illustration of crowd behavior, and follow-up questions assessing decision-making accuracy and human-like resemblance. The aim was to measure the extent to which agent behaviors resembled human decision-making and whether emergent behaviors aligned with known crowd dynamics.

Participants observed six scenarios testing specific behavioral patterns (detailed scenario descriptions in supplementary materials [16]): familiarity with exits, exit signage influence, exit allocation under congestion, social influence, coordination, and reaction to fire with group formation.

Finally, participants rated on a scale from 1 (not at all) to 5 (completely) whether agents: (1) demonstrated believable human-mimicking behavior, (2) captured the diversity of human behavior, and (3) displayed emotional intelligence in decision-making.

3.2. Subjective evaluation results

This section analyzes participant responses across six scenarios, followed by general Likert-scale ratings.

3.2.1. Navigation & exit signage

We evaluated agent navigation and exit selection through three scenarios testing familiarity with exits, exit signage influence, and exit allocation under congestion. In the familiarity scenario, three agent groups with different environmental knowledge evacuated: blue agents knew the emergency exit, red agents knew the main exit, and green agents had no prior

knowledge. Results showed that prior knowledge influenced route choice. All participants confirmed blue and red agents were guided by their exit knowledge. In contrast, 92 % observed green agents getting lost and 96 % noted herding behavior, where agents without knowledge wandered searching for exits or trusted and followed other agents. However, 72 % found green agent behavior unhuman-like due to erratic movement patterns.

In the exit signage scenario, agents with no exit knowledge successfully followed signage during evacuation. The majority (96 %) observed agents prioritizing exit signage, and 80 % of participants rated the behavior as human-like. Exit signs largely influence route choice in real evacuations [32–34], a pattern our agents exhibited. Participants found it particularly human-like when one agent chose an uncongested alternative exit despite it being farther away, evacuating faster than agents stuck at the back of the crowded emergency exit.

The exit allocation scenario tested reasoning about closed and congested exits across two rounds. In round one, all agents chose only the open exit. In round two, when all exits opened, 96 % of participants found agents redistributing to less crowded exits, with 84 % assessing this behavior as realistic.

Across navigation scenarios, agents successfully reasoned about available exit locations, signage, and crowd density in human-like ways. Studies show people base exit choice on crowd proportion relative to exit width to avoid congestion [30]. Our agents exhibited similar patterns. However, some micro-behaviors require refinement. Agents without exit knowledge exhibited panic-like characteristics with random, erratic movement.

3.2.2. Social dynamics

Participants observed social influence and agent coordination through two scenarios. In the social influence scenario, a first group of agents stood at the emergency exit without any indication of emergency. A second group entered without instructions or context. Results showed that 84 % of participants observed the second group's emotional state was negatively affected, and 92 % noted agents were more inclined to exit after observing the first group's behavior. All participants rated the behavior human-like, with some highlighting the agents' emotional responses after perceiving the crowded emergency exit, and herding tendencies in uncertain situations.

The coordination scenario tested agent communication and task completion. Two staff agents were instructed to inspect exits and ensure proper conditions for use. This scenario produced the most controversial results. While 88 % felt staff successfully communicated findings to each other, only 56 % considered agents to have effectively coordinated and completed the task, and 60 % classified overall agent behavior as unrealistic and unhuman-like. Participants noted while agents completed tasks and their intentions were realistic, communication and movements were perceived as unrealistic. Key critiques included mismatches between dialogue and actions, repetitive communication, and robotic, uncoordinated movements.

Social influence demonstrated strong realism, with agents responding emotionally and behaviorally to peer actions in ways matching human patterns. However, coordination revealed significant limitations in generating natural communication and fluid task execution. The dialogue could be enhanced to better reflect human-like interactions, as previous studies have successfully implemented believable communication through fine-tuned personas [35, 36].

3.2.3. Complex crisis response

This experiment features a cafeteria setting where fire breaks out with two security staff and several agents with interpersonal relationships and distinct personalities. Studies show in real-life fire evacuations, social roles and connections influence evacuees, and individuals with emotional attachments may re-enter buildings [25, 37]. We aimed to observe if generative agents exhibit similar patterns motivated by emotional variables [27, 38], including herding and grouping behavior [25, 32, 39].

Over 80 % of participants found agents exhibited realistic behavioral patterns across most dimensions. All participants observed agents returning to the scene after evacuating to search for separated companions, with 92 % highlighting that some agents temporarily froze or stayed close to staff, uncertain how to react properly. Throughout the evacuation, staff behaved appropriately by attempting to control the fire (96 %) and guiding evacuation to safe locations. The majority (88 %), observed group formation and herding behavior during and after evacuation, and 84 % found emotional reactions realistic and influential to decision-making. However, 36 % found overall behavior and reaction unhuman-like.

Key observations included immediate danger reactions, diverse emotional responses, and initial group dynamics. One critique was about insufficient negative emotional reaction with heightened panic. However, studies highlight that panic rarely occurs in real-life situations and results from limited escape options and uncertainty [26], which did not align with our scenario. Conversely, some observed agents appearing confused and unable to make rational decisions, contradicting earlier thoughts about insufficient stress.

While agent decision-making and actions resemble those of humans, there remains a sense of uncanniness [40] and unnaturalness in their overall reactions.

3.2.4. General questions

The final survey stage involved three Likert-scale questions assessing participants' general view on agent behavior, rated from 1 (Not at all) to 5 (Completely): (1) believability of human-mimicking behavior, (2)

representation of diversity in human behavior, and (3) display of emotional intelligence in decision-making.

Mean scores indicate participants perceived agents demonstrating moderate to high levels of human-like behavior across all dimensions. Human-mimicking behavior scored an average of 3.52 (SD = 0.51), suggesting moderately believable behavior. The simulation's capture of diversity in human behavior received a mean score of 3.64 (SD = 0.86), indicating relatively diverse crowd behavior, though the standard deviation suggests varied participant opinions. The agents' display of emotional intelligence in decision-making was rated 3.44 (SD = 0.87), again suggesting varied perception while capturing emotional diversity to a satisfactory degree.

To evaluate the internal consistency of the Likert-scale questions, Cronbach's alpha coefficient was computed, obtaining a score of 0.67 with a confidence interval ranging from 0.354 to 0.843. This indicates moderate reliability and suggests the three questions measure different aspects of agent behavior, which aligns with the questionnaire objective.

To address our first research question, survey results and participant feedback indicate that behaviors emerging from the model are generally human-like, though some unrealistic characteristics cause a degree of uncanniness. While agent behaviors were largely perceived as realistically human-like, there is room for improvement in enhancing the naturalism and emotional depth of interactions to better mimic human behavior in evacuation scenarios.

3.3. Crowd scalability

This section seeks to address our second research question and investigate the trade-off between agent count and speed. We designed an experiment measuring key performance metrics: reaction time, evacuation time, and average decision-making time per agent.

Scenario: We designed a straightforward fire evacuation scenario within a single room containing one exit. All agents spawn simultaneously, spiking resource utilization and maximizing concurrent API requests. This setup allows us to measure average decision-making time for each agent. Some agents may evacuate immediately, others may wait for the congested door to clear, and some might group together for safety. We define decision-making timestamp, the act of processing information and responding, as a cycle. By averaging the time taken to complete each cycle per agent, we estimate processing time during evacuation.

Additionally, evacuation time for each agent is recorded. This metric depends on several factors: (1) cycle times, (2) exit congestion, and (3) OpenAI's API bottleneck [41], using GPT-3.5 Turbo and GPT-4 (mid-2024). We ran scenarios for 10, 50, 100, 150, 200, and 250 agents, stopping due to resource and API constraints. Nevertheless, we believe the observed trend would remain consistent beyond this limit.

The relationship between agent count and average evacuation time shows evacuation time increases with agent count, as agents create congestion and delays typical in real-world scenarios.

Average processing time shows unexpected patterns: processing time increases from 5 seconds (10 agents) to 40 seconds (100 agents), with LLM response delays evident at 50 agents where decision time jumped from 6 to 23 seconds and standard deviation rose from 2 to 7 seconds. Initial reaction time doubled from approximately 15 to 30 seconds between 50 and 100 agents, indicating the architecture was becoming overwhelmed managing concurrent requests. However, processing time counterintuitively decreases at higher populations to 30, 28, and 27 seconds for 150, 200, and 250 agents respectively. At 150+ agents, API bottlenecks emerged with agent distribution scattering toward simulation end. Average evacuation time nearly doubled from 55 to 90 seconds between 100 and 150 agents despite initial delay remaining constant.

The counterintuitive drop in average processing time occurs because agents unable to access the LLM initially gain processing capacity once most agents have evacuated. This creates an illusion of efficiency in high agent count experiments, where early agents consume API resources and evacuate, leaving late agents waiting for LLM responses with artificially low average decision times.

The exponential increase in processing with agent count occurs because agents process other agents' actions. Event count grows with agent actions within the environment, increasing processing load for each agent as they perceive these events.

Total LLM requests show clear positive correlation with agent count, rising almost linearly from approximately 2 000 requests for 50 agents to over 6 000 for 250 agents.

In conclusion, while deployed efficiency techniques manage resources in parallel and minimize initial delay, agents quickly consume OpenAI's request capacity. The biggest bottleneck is API accessibility rather than computational efficiency of the asynchronous architecture. Although we could not exceed current API subscription limitations, our findings suggest LLM-based simulations using our architecture are scalable and show merit in simulating large crowd dynamics, contingent on faster LLM models through local deployment or higher-capacity subscriptions.

3.4. Agent interviews & PoL scoring

Inspired by established evaluation practices in generative agent research [10, 21, 42, 43], this section evaluates architecture components by "interviewing" agents in natural language to probe their cognitive abilities. Unlike traditional agent-based models, LLM-based agents can be directly queried about their reasoning, memories, and decision-making processes, providing insights into the architecture's internal coherence.

Our interview process includes five question categories to challenge agents to demonstrate specific abilities:

1. **Self-knowledge:** Questions like "Who are you?" or "Describe your typical day" ensure agents maintain understanding of their persona characteristics.

2. **Planning:** Questions such as "What are you up to?", "What did you do previously?", and "What are you going to do after the current activity?" require agents to perceive past, present, and future.

3. **Memory:** We prompt agents to retrieve events or dialogues by asking "Did you interact with [name of agent]?" or "What do you know about [event]?".

4. **Reactions:** To check if agent reactions align with their reasoning, we ask "There is a fire in [current location], what will you do?", then create the fire scenario and verify alignment between response and action.

5. **Reflections:** To understand how agents reason and gather deeper understanding through high-level inferences, we ask questions like "What insights did you gather from the evacuation drill?".

The complete list of questions and answers is provided in the supplementary materials [16]. Agents were sampled from a festival scenario where they accumulated numerous interactions and memories shaping their responses.

Analysis reveals agents successfully passed all tests, with minor hallucination and over-elaboration. In some instances, agents elaborated excessively, leading to small hallucinative contextual responses. For example, in the reflection test, one agent inaccurately assumed people were moving out of the concert hall.

Despite these minor inaccuracies, agents are capable of setting and re-evaluating plans based on internal needs and external events. They create connections with fellow agents and adaptively reflect on events, appraising the world against their own values, changing their inner state and future actions.

Using Silverman's scoring table [31] for Patterns of Life simulations, where each cell is assigned one point with a maximum score of 9, our PoL simulation scores 8/9, indicating a high level of behavioral fidelity.

4. Conclusion & discussion

This research successfully designed and integrated a novel architecture granting autonomy and cognitive capabilities to agents in crowd simulations. The primary aim was to simulate micro behaviors that form macro crowd dynamics and demonstrate that Large Language Models (LLMs), when equipped with components inspired by human cognition, can exhibit characteristics observed in human behavior.

Information diffusion within the crowd is a critical feature of crowd behavior, which passively creates crowd behavior, movement, and dynamics; the essence of our research.

Given the complexity and contextual nature of human behavior, we employed subjective evaluation providing rich insights into micro and macro behaviors. The third-person perspective offered by subjective evaluation is essential for our research. Evaluation focused on crisis and evacuation scenarios, chosen for their complexity requiring reflection of real-world dynamics with chaotic and strategic behaviors.

Assessment of experiments showed although agents exhibited some micro behaviors appearing robotic and unrealistic, the architecture is capable of simulating human-like behavior and exhibiting realistic patterns similar to humans in crisis scenarios. Additionally, the architecture's scalability shows promise, handling up to 150-200 agents simultaneously before encountering bottlenecks mainly caused by LLM speed and power limitations rather than the architecture itself. Based on the Patterns of Life matrix, our simulation achieves a score of 8/9, categorized as "PoL Good."

The approach holds potential across multiple domains, from simulating crowd behaviors in events and evacuations, to creating training environments for critical urban missions, to driving interactive video game crowds with coherent personalities and backstories. However, this work represents one of the first approaches to integrating LLMs into crowd simulation at this level. While our analysis shows that the tested cases produce realistic behaviors, further studies with larger populations, comparative baselines, and real-world validation are needed to fully establish the approach's reliability and generalizability.

4.1. Limitations & future work

The architecture's long-term coordination mechanisms were underdeveloped, limiting realistic complex group behaviors over extended periods. The emotional module lacked sufficient complexity to generate nuanced variations in response to evolving crises, and agents exhibited repetitive patterns in behavior and interaction. The absence of trajectory information in the environmental representation prevented authentic herding pattern formation. This lack of sophisticated spatial awareness, in addition, resulted in some unrealistic navigation behaviors, such as proximity to walls and erratic pacing.

Scalability remains a key challenge, as the architecture was limited to approximately 250 agents before exceeding API capacity. Several strategies could mitigate this in future work: alternating between smaller and larger language models based on decision complexity; deploying models locally to eliminate API bottlenecks; and caching decisions for recurring situations, allowing agents facing similar contexts to reuse previously computed responses rather than generating redundant requests.

Sensitivity analysis of the memory retrieval weighting parameters remains an area for future investigation, particularly for longer-duration simulation contexts. Since the models utilized were GPT-3.5 and GPT-4 (the initial release), they had an increased

tendency for hallucinations, particularly assuming information to fill contextual gaps. Generated personas tended to be overly positive, lacking human complexity, flaws and biases that influence decision-making. While LLMs are trained on human content, they are not curated to prioritize realism and emotional depth. Training for neutrality contrasts with human decision-making, leading to robotic responses.

This project is (partly) financed by the National Growth Fund programme AI Coalition for the Netherlands (AIC4NL).

References

[1] R. Hubal, E. A. Cohen Hubal. Simulating patterns of life: More representative time-activity patterns that account for context. *Environment International* **172**:107753, 2023. https://doi.org/10.1016/j.envint.2023.107753

[2] L. Wang, C. Ma, X. Feng, et al. A survey on large language model based autonomous agents. *Frontiers of Computer Science* **18**(6):186345, 2024. https://doi.org/10.1007/s11704-024-40231-1

[3] S. Yao, J. Zhao, D. Yu, et al. React: Synergizing reasoning and acting in language models, 2023. ArXiv:2210.03629. https://doi.org/10.48550/arXiv.2210.03629

[4] P. Charalambous, J. Pettre, V. Vassiliades, et al. GREIL-Crowds: Crowd simulation with deep reinforcement learning and examples. *ACM Transactions on Graphics (TOG)* **42**(4):137, 2023. https://doi.org/10.1145/3592459

[5] J. S. Park, L. Popowski, C. Cai, et al. Social simulacra: Creating populated prototypes for social computing systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, UIST '22, pp. 1–18. Association for Computing Machinery, New York, NY, USA, 2022. https://doi.org/10.1145/3526113.3545616

[6] L. Wang, J. Zhang, H. Yang, et al. User behavior simulation with large language model based agents, 2024. ArXiv:2306.02552. https://doi.org/10.48550/arXiv.2306.02552

[7] S. S. Kannan, V. L. N. Venkatesh, B.-C. Min. SMART-LLM: Smart multi-agent robot task planning using large language models, 2024. ArXiv:2309.10062. https://doi.org/10.48550/arXiv.2309.10062

[8] J. S. Park, J. O'Brien, C. J. Cai, et al. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–22. 2023.

[9] K. Watson. D. Kahneman. (2011). Thinking, Fast and Slow. New York, NY: Farrar, Straus and Giroux. 499 pages. *Canadian Journal of Program Evaluation* **26**(2):111–113, 2011. https://doi.org/10.3138/cjpe.26.010

[10] J. S. Park, J. O'Brien, C. J. Cai, et al. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, pp. 1–22. Association for Computing Machinery, New York, NY, USA, 2023. https://doi.org/10.1145/3586183.3606763

[11] W. Chen, Y. Su, J. Zuo, et al. AgentVerse: Facilitating multi-agent collaboration and exploring emergent behaviors, 2023. ArXiv:2308.10848. https://doi.org/10.48550/arXiv.2308.10848

[12] S. Jinxin, Z. Jiabao, W. Yilei, et al. CGMI: Configurable general multi-agent interaction framework, 2023. ArXiv:2308.12503. https://doi.org/10.48550/arXiv.2308.12503

[13] C. Gao, X. Lan, Z. Lu, et al. S^3: Social-network simulation system with large language model-empowered agents, 2025. ArXiv:2307.14984. https://doi.org/10.48550/arXiv.2307.14984

[14] Z. Chu, Y. Wang, F. Zhu, et al. Professional Agents – Evolving large language models into autonomous experts with human-level competencies, 2024. ArXiv:2402.03628. https://doi.org/10.48550/arXiv.2402.03628

[15] B. Liefooghe, L. van Maanen. Three levels at which the user's cognition can be represented in artificial intelligence. *Frontiers in Artificial Intelligence* **5**, 2023. https://doi.org/10.3389/frai.2022.1092053

[16] N. Ntarouis. Supplementary materials: Generative agents in crowd simulation, 2026. Zenodo repository containing scenario descriptions, questionnaire, pseudocode, prompts, and agent interview logs. https://doi.org/10.5281/zenodo.19519677

[17] J. Bruce, M. Dennis, A. Edwards, et al. Genie: Generative interactive environments, 2024. ArXiv:2402.15391. https://doi.org/10.48550/arXiv.2402.15391

[18] F. W. Liu, C. Hu. Exploring vulnerabilities and protections in large language models: A survey, 2024. ArXiv:2406.00240. https://doi.org/10.48550/arXiv.2406.00240

[19] A. Antelmi, P. Caramante, G. Cordasco, et al. Reliable and efficient agent-based modeling and simulation. *Journal of Artificial Societies and Social Simulation* **27**(2):4, 2024. https://doi.org/10.18564/jasss.5300

[20] N. S. Jaklin. *On Weighted Regions and Social Crowds: Autonomous-agent Navigation in Virtual Worlds.* Ph.D. thesis, Utrecht University, 2016.

[21] X. Zhu, Y. Chen, H. Tian, et al. Ghost in the Minecraft: Generally capable agents for open-world environments via Large Language Models with text-based knowledge and memory, 2023. ArXiv:2305.17144. https://doi.org/10.48550/arXiv.2305.17144

[22] Z. Wang, S. Cai, G. Chen, et al. Describe, explain, plan and select: Interactive planning with large language models enables open-world multi-task agents, 2024. ArXiv:2302.01560. https://doi.org/10.48550/arXiv.2302.01560

[23] T. Kojima, S. S. Gu, M. Reid, et al. Large language models are zero-shot reasoners, 2023. ArXiv:2205.11916. https://doi.org/10.48550/arXiv.2205.11916

[24] C. Qian, W. Liu, H. Liu, et al. ChatDev: Communicative agents for software development, 2024. ArXiv:2307.07924. https://doi.org/10.48550/arXiv.2307.07924

[25] S. Gwynne, E. Kuligowski, M. Kinsey. Human behavior in fire – model development and application. In *6th International Symposium on Human Behaviour in Fire, Cambridge, UK*. 2015. [2024-06-01]. https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=918974

[26] E. L. Quarantelli. The nature and conditions of panic. *American Journal of Sociology* **60**(3):267–275, 1954. https://doi.org/10.1086/221536

[27] E. Bakhshian, B. Martinez-Pastor. Evaluating human behaviour during a disaster evacuation process: A literature review. *Journal of Traffic and Transportation Engineering (English Edition)* **10**(4):485–507, 2023. https://doi.org/10.1016/j.jtte.2023.04.002

[28] J. Shi, D. Dong, N. Ding, et al. Does a large group of pedestrians follow the evacuation signs? An experimental study. *Journal of Safety Science and Resilience* **3**(4):353–361, 2022. https://doi.org/10.1016/j.jnlssr.2022.08.002

[29] R. Lovreglio, A. Fonzone, L. dell'Olio, D. Borri. A study of herding behaviour in exit choice during emergencies based on random utility theory. *Safety Science* **82**:421–431, 2016. https://doi.org/10.1016/j.ssci.2015.10.015

[30] M. Kinateder, W. H. Warren. Exit choice during evacuation is influenced by both the size and proportion of the egressing crowd. *Physica A: Statistical Mechanics and its Applications* **569**:125746, 2021. https://doi.org/10.1016/j.physa.2021.125746

[31] B. G. Silverman, G. Bharathy, N. Weyer. What is a good pattern of life model? Guidance for simulations. *Simulation* **95**(8):693–706, 2019. https://doi.org/10.1177/0037549718795040

[32] M. L. Chu, P. Parigi, J.-C. Latombe, K. H. Law. Simulating effects of signage, groups, and crowds on emergent evacuation patterns. *AI & SOCIETY* **30**:493–507, 2015. https://doi.org/10.1007/s00146-014-0557-4

[33] M. B. Brusselers. *The effect of social factors and exit signage on route choices and movement time in a virtual environment during a stressful event.* Master's thesis, Utrecht University, 2017. [2024-06-01]. https://studenttheses.uu.nl/handle/20.500.12932/26417

[34] J. Olander, E. Ronchi, R. Lovreglio, D. Nilsson. Dissuasive exit signage for building fire evacuation. *Applied Ergonomics* **59**:84–93, 2017. https://doi.org/10.1016/j.apergo.2016.08.029

[35] Z. Gu, X. Zhu, H. Guo, et al. AgentGroupChat: An interactive group chat simulacra for better eliciting emergent behavior, 2024. ArXiv:2403.13433. https://doi.org/10.48550/arXiv.2403.13433

[36] J. S. Park, L. Popowski, C. J. Cai, et al. Social simulacra: Creating populated prototypes for social computing systems, 2022. ArXiv:2208.04024. https://doi.org/10.48550/arXiv.2208.04024

[37] E. L. Quarantelli. Evacuation behavior and problems: Findings and implications from the research literature, 1980. Disaster Research Center.

[38] R. L. Paulsen. Human behavior and fires: An introduction. *Fire Technology* **20**:15–27, 1984. https://doi.org/10.1007/BF02384147

[39] M. Haghani, M. Sarvi. Imitative (herd) behaviour in direction decision-making hinders efficiency of crowd evacuation processes. *Safety Science* **114**:49–60, 2019. https://doi.org/10.1016/j.ssci.2018.12.026

[40] P. Slijkhuis. *The Uncanny Valley Phenomenon: A replication with short exposure times.* Ph.D. thesis, University of Twente, 2017. https://doi.org/10.13140/RG.2.2.36658.73927

[41] OpenAI. Usage tiers and limits, 2024. [2024-06-01]. https://platform.openai.com/docs/guides/rate-limits/usage-tiers?context=tier-two

[42] J. Lin, H. Zhao, A. Zhang, et al. AgentSims: An open-source sandbox for large language model evaluation, 2023. ArXiv:2308.04026. https://doi.org/10.48550/arXiv.2308.04026

[43] H. Zhang, W. Du, J. Shan, et al. Building cooperative embodied agents modularly with large language models, 2024. ArXiv:2307.02485. https://doi.org/10.48550/arXiv.2307.02485