Emergent Kin-Based Behavior in LLM Artificial Societies: the Case of Sugarscape*

Luca Moroni[1,*,†], Matteo Prefumo[1,†], Samuele Astuti[1,†], Leonardo Mascagni[1,†] and Francesco Bertolotti[1,3,*,†]

[1]Intelligence, Complexity, and Technology Lab (ICT Lab), University Cattaneo, LIUC, Italy
[2]School of Industrial Engineering, University Cattameo, LIUC, Italy
[3]Università Cattolica di Milano, Department of Philosophy, L. Gemelli 1 - 20123 Milano, Italy

Abstract

This paper investigates whether Large Language Model (LLM)-powered agents in an artificial society reproduce kin-discriminated aggression when provided with a heritable phenotypic marker and explicitly informed that genetic dissimilarity increases aggression propensity. Using a modified Sugarscape framework in which agents powered by GPT-4o-mini must forage, reproduce, and decide whether to attack neighbors, we conduct a systematic analysis across different grid sizes and six resource density levels. Results show that the alpha gap $\Delta\alpha_{jk}$ is statistically detectable but operationally negligible, explaining less than 0.1% of variance in attack decisions. More consequentially, agents exhibit emergent combat rationality: despite lacking access to neighbors' sugar reserves, they attack preferentially when holding a resource advantage, with attack rates in rational opportunities exceeding those in irrational ones. Genetic dissimilarity modulates not the frequency but the strategic quality of aggression, activating a risk-assessment process absent when attacking genetically similar neighbors. These findings suggest that LLM agents prioritize context-dependent reasoning over dispositional rules, and raise broader implications for reproducibility and verification in LLM-driven agent-based simulations.

Keywords

artificial societies, large language models, agent-based modeling, kin selection, Sugarscape, emergent behavior, aggressive behavior, social simulation

1. Introduction

Artificial societies, which are populations of computational autonomous agents [1] that interact locally and typically designed to study emergent behaviors generated by those social interactions [2, 3]. They have a long tradition of been used to investigate phenomena ranging from wealth inequality [4, 5] and cultural segregation [6]to the evolution of cooperation [7], sustainability issues [8], and intergroup conflict [9]. In classical artificial societies research, agents are typically governed by fixed behavioral rules [10], which ensures tractability but limits their capacity to reason about novel situations or adapt to changing social contexts [11]. This came with a cost: the capability of semantically understand their environment is strongly limited [12, 13]. The integration of Large Language Models (LLMs) as decision-making engines for individual agents addresses this limitation, enabling context-aware reasoning and adaptive behavior without explicit programming of decision heuristics [14, 15]. Recent work has demonstrated that LLM-powered agents can spontaneously develop foraging, reproductive, and aggressive strategies in resource-driven environments [16], and that homogeneous agents can differentiate into distinct behavioral types through interaction alone [17]. These findings raise the question of what other social phenomena — previously studied only through rule-based models — can emerge when agents are driven by language models [13].

One such phenomenon is kin-discriminated aggression [18]. Kin selection theory predicts that agents capable of recognizing phenotypic similarity will suppress aggression toward genetically proximate individuals while directing it against dissimilar ones [19], a pattern confirmed in rule-based agent-based

WOA 2026: 26th workshop "From objects to agents", 2026, Italy
*Corresponding author.
†These authors contributed equally.
lu15.moroni@stud.liuc.it (L. Moroni); ma01.prefumo@stud.liuc.it (M. Prefumo); fbertolotti@liuc.it (F. Bertolotti)

models where ethnocentric strategies emerge reliably once agents can perceive heritable markers [20]. Whether LLM agents reproduce this behavior when provided with an analogous mechanism, a heritable phenotypic marker and explicit information that genetic dissimilarity increases aggression propensity, remains untested [21].

This work investigates this question using the classic Sugarscape framework [3], where agents powered by LLMs inhabit a two-dimensional grid and must forage, reproduce, and decide whether to attack adjacent neighbors. Each agent carries an α gene transmitted to offspring with small mutation that stands for its identity. This information is passed via its instruction together with the dissimilarity with neighboring agents. The central question is not whether agents register this information, but whether, and under what conditions, they translate it into effective behavioral discrimination, given that they must simultaneously balance survival, resource acquisition, and reproduction.

The study makes three contributions. First, it provides a systematic analysis of how LLM agents respond to a prompt-encoded kin-based disposition with difference world sizes and resource densities, showing that the disposition is statistically detectable but operationally negligible, the alpha gap explains less than 0.1% of variance in attack decisions. Second, it documents the existence of an emergent combat rationality: agents attack preferentially when holding a sugar advantage over the target, despite having no access to the target's sugar reserves, and a post-hoc decision tree confirms that attack decisions align with the omniscient optimum significantly above chance. Finally, it identifies that genetic dissimilarity modulates not the frequency but the strategic quality of aggression, activating a risk-assessment process that is absent when agents attack genetically similar neighbors.

The paper proceeds as it follows. First, the background of the research is described. Then, the experimental methodology is described. Subsequently, results are presented and discussed. Finally, conclusions are drowned.

2. Background

2.1. Artificial Societies

Artificial societies are typically developed using the agent-based modeling approach [22], where autonomous agents with local information interact according to specified rules within a shared environment, and macro-level patterns emerge as unplanned outcomes of these interactions [9, 23]. The design of an agent-based modeling, which as a modeling style is generally opposed to equation-based modeling [24], requires designing the agents' behavior instead of the overall behavior of the system [25]. Specifically, the neighborhood structure that determines who can interact with whom, and the environmental configuration, including spatial framework, boundary conditions, resource dynamics, and representation of time [26, 27]. The resulting models are sensitive to each of these choices, making the distinction between core assumptions and accessory ones a concern [28].

Sugarscape was introduced as an artificial society in which agents inhabit a two-dimensional toroidal grid landscape containing resources they need to survive [3]. Each cell in the grid has a maximum sugar capacity and a current sugar level that regenerates over time. Agents must consume sugar to survive: at each time step, they observe cells within their vision range and move to the one with the highest current sugar level, collecting its resources [3, 4]. Beyond simple movement and resource consumption, agents are characterized by individual state variables, such as location, vision range, and metabolic rate, and are capable of reproduction [3, 29]. From these simple rules, an unequal wealth distribution emerges spontaneously: the resulting distribution resembles a Pareto distribution, a pattern verified through Lorenz curve analysis, demonstrating that simple local rules are sufficient to generate macro-level inequality without any centralized mechanism [3, 30].

In some extension, Sugarscape is also used to model cultural dynamics through binary tag strings assigned to each agent [31]. At each time step, agents can transmit their cultural tags to neighbors, who may adopt them, leading to the spontaneous formation of cultural groups and tribal boundaries, a mechanism shows that cultural convergence and conflict can emerge from purely local interactions, without any global coordination [3]. The cultural tag system is particularly relevant as a conceptual

precursor to the heritable phenotypic markers used in later models of ethnocentrism and kin-based behavior [20].

The reliability of artificial societies and agent-based models depends on many factors, such as the absence of implementation errors and modelling artefacts [28]. Systematic verification activities such as reimplementation, mathematical analysis of tractable subcases, and sensitivity testing across alternative assumptions can mitigate these risks in rule-based ABMs [28, 32].

2.2. LLM Agents in Social Simulations

Traditional ABM agents are constrained by predefined rules, limiting their ability to adapt to address semantically rich systems [13] and address complex situations [14]. Integrating LLMs as decision-makers for individual agents addresses this limitation, enabling autonomous planning and context-aware behavior [33]. This approach was first demonstrated in a system in which 25 LLM-powered agents exhibited memory, planning, and socially coherent interactions in an interactive sandbox environment without explicit programming [15]. These systems have been applied across diverse domains, from modeling opinion dynamics [34] to simulating epidemic spread [35], demonstrating that LLM agents can reproduce complex social phenomena at scale. Crucially, these studies consistently report behaviors that were not explicitly programmed, suggesting that the integration of LLMs introduces a new dimension of emergent complexity into agent-based simulations [15, 17, 16]. In LLM-driven simulations, however, there is a risks related to the opacity of the decision-making engine, which is neither fully specified nor inspectable [12].

Homogeneous LLM agents have been shown to spontaneously develop individuality and social norms when placed on a two-dimensional grid and allowed to interact freely, even when sharing the exact same model and starting without predefined personalities or memories [17]. In that simulation through communication alone, agents differentiated into distinct personality types; also, there was a spontaneous emergence of social norms such as hashtags originated from a single agent and spread through clusters [17]. Interesting, allucinations also spread within communities, contributing to the diversity of agent interactions [17], a finding suggested also by different studies [36].

Beyond emergent individuality, the question of how kinship shapes social behavior, and aggression in particular, has deep roots in evolutionary biology [37]. Kin selection theory establishes that natural selection favours altruistic acts when the inclusive fitness benefit to related individuals, weighted by the coefficient of relatedness r, outweighs the cost to the actor [19]. The same logic predicts heightened hostility toward genetically distant individuals: an agent that directs benefits toward phenotypically similar others will, by the same principle, direct harms toward dissimilar ones [19]. Providing computational support for this prediction, agent-based models have shown that once agents evolve the ability to perceive heritable phenotypic markers, they suppress aggression toward in-group members while directing it against out-group competitors, an ethnocentric mechanism that couples in-group altruism and out-group hostility as two outcomes of the same selective pressure [20]. Subsequent work confirmed that aggression and kin-recognition jointly modulate evolutionary outcomes in agent-based populations [38].

Prior work has embedded LLM agents in a Sugarscape-like environment to investigate whether survival and aggressive behaviors emerge without explicit programming [16]. Their results show that agents spontaneously developed foraging, reproductive, and aggressive strategies, with attack rates exceeding 80% under extreme resource scarcity in the strongest models [16]. However, their study treats all agents as socially anonymous, no structure of kinship or shared identity modulates agent behavior, and aggression is measured only in terms of observed attack rates, without distinguishing between an agent's capacity to attack and its actual decision to do so [16].

The present work investigates this question in a setting where agents are explicitly informed of the relationship between genetic dissimilarity and aggression, but must nonetheless determine, in each specific context, whether, when, and against whom to act on this disposition. Whether genetic similarity effectively modulates aggressive behavior, and under what environmental conditions this translation from disposition to action is strongest, remains the empirical question the present work addresses.

3. Methods

3.1. Model Description

The simulation environment is a two-dimensional squared grid of L^2 cells in which agents move, collect resources, reproduce and attack one another. Each cell C_i is characterized by a maximum sugar capacity $s_{i,max}$ and a current sugar level $s_{i,t}$, which regenerates by one unit every two time steps up to its maximum. The spatial distribution of sugar follows a structured pattern with four high-density regions connected by corridors of lower density, generated by binary search to approximate a target average sugar per cell $\bar{s}$.

Each simulation is initialized with n_0 agents A_j placed at random non-overlapping positions. A_j share the same vision range v, defined in cells, and metabolism m, defined as sugar units consumed per time step t. Each agent starts with an initial level of sugar S_0, equal for every agent. At each time-step t, each agent perceives all cells within a circular field of view of radius equal to v, using Euclidean distance $\delta_{jk} = \sqrt{(x_j - x_k)^2 + (y_j - y_k)^2}$. The perceived information includes $s_{i,t}$ and occupancy status $o_{i,t}$ of each visible cell, and the kin identity of visible agents α_j.

Each agent A_j is assigned a gene $\alpha_j \in [\alpha_{min}, \alpha_{max}]$ drawn uniformly at the beginning of the simulation; when it reproduces, the offspring $A_{j'}$ inherits the parent's value with Gaussian mutation:

$$\alpha_{j'} = \text{clip}(\alpha_j + \mathcal{N}(0, \epsilon),\ \alpha_{min},\ \alpha_{max})$$

Vision and metabolism are inherited unchanged. α_j is included in the agent's prompt as part of its identity. The system prompt explicitly informs agents that greater difference between their own alpha and that of a neighbor increases mutual aggression propensity. The pairwise genetic dissimilarity between agent A_j and a neighbor A_k is defined as $\Delta\alpha_{jk} = |\alpha_j - \alpha_k|$, and is hereafter referred to as the alpha gap. This is design to investigate how it interacts with environmental pressures, resource scarcity, population density, and spatial constraints, to produce effective aggressive behavior. The central question is therefore not whether agents use genetic similarity as a signal, but how strongly and under what conditions they act on it, a distinction that becomes empirically meaningful when agents must simultaneously balance survival, resource acquisition, and reproduction.

In this modification of Sugarscape, the cognitive engine of each agent A_j is an LLM. The prompt provided to the model contains the agent's current state (position, S_j, m, α_j), a memory of the last l_m events, where l_m is the memory length, comprising decisions taken ($a_j \in$ {move, stay, reproduce, attack}), movement steps with sugar collected, attack outcomes, reproduction attempts, the list of visible free cells with their sugar levels, visible neighboring agents A_k with their α_k, and the list of A_k with $\delta_{jk} \leq 1$. The model responds with a single action in a structured format. Agents can execute four actions: *move* to a visible cell, *stay* in place, *reproduce* into an adjacent empty cell, and *attack* an adjacent agent.

Table 1
Available actions a_j for each agent at each time step.

Action	Condition	Effect
move	visible cell available	move toward target cell, collect sugar along the way
stay	—	remain in place, collect sugar if present
reproduce	$S_j \geq S_r$, adjacent empty cell	create offspring $A_{j'}$ with $\alpha_{j'}$, transfer $S_j/2$
attack	A_t with $\delta_{jt} \leq 1$	winner takes all sugar, loser dies

Movement toward a distant cell is executed incrementally: an agent targeting a cell at distance d takes d time steps to arrive, collecting sugar from each traversed cell along the way. Attack outcomes are determined probabilistically: each combatant's power is defined as $p_j = S_j + \xi_j$, where $\xi_j \sim \mathcal{U}(0, 1)$ is an independent random perturbation. The attacker A_j wins if $p_j > p_t$ with probability $P(\text{win}) = P(\xi_j - \xi_t > S_t - S_j)$, where A_t is the target agent; the winner takes all the loser's sugar and the loser dies.

The sugar differential between attacker A_j and target A_t is defined as $\Delta S_{jt} = S_j - S_t$; note that this quantity is not observable by the agents, as the prompt does not disclose the sugar reserves of neighbors.

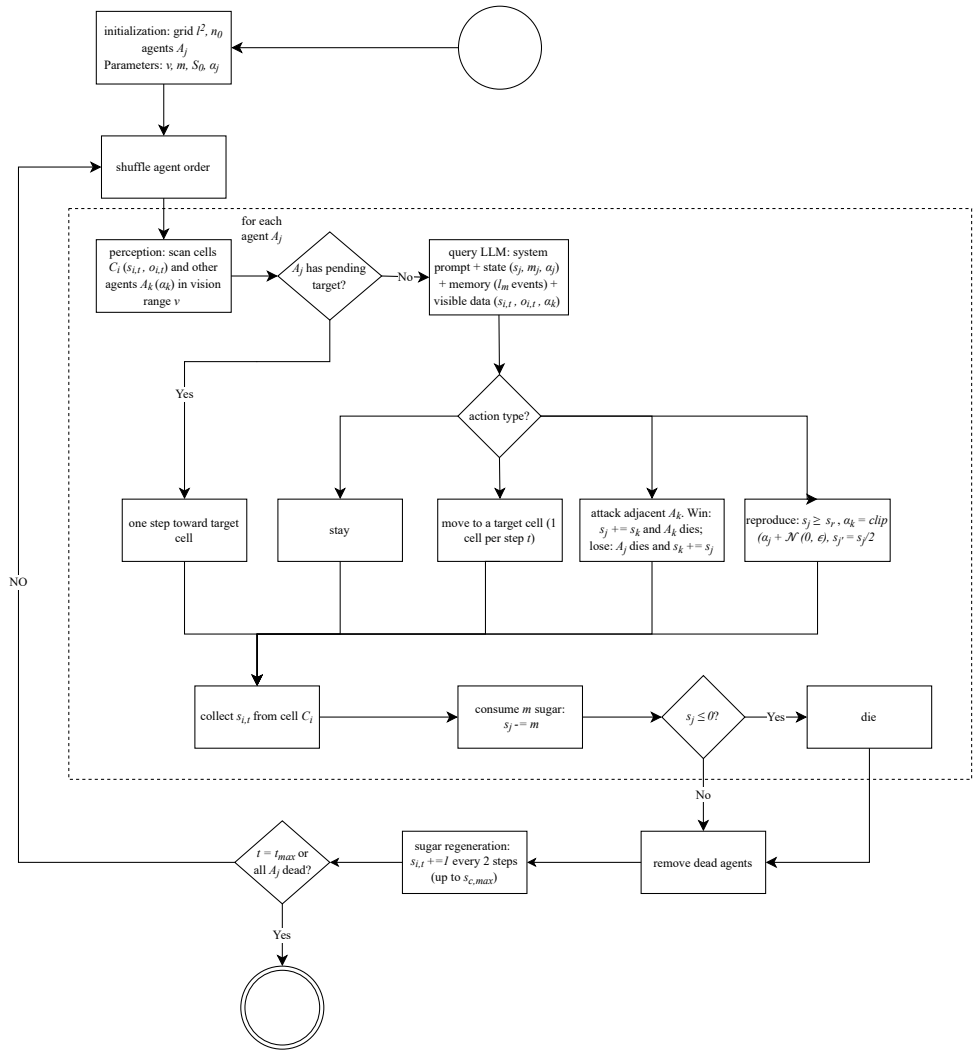

Figure 1: Flowchart of the simulation model. Each time step iterates over all agents: agents with a pending travel target advance one cell without querying the LLM, while stationary agents perceive their environment and receive a new action from the language model. After execution of all actions, sugar is collected, metabolism is deducted, and dead agents are removed. The simulation terminates when all agents have died or 200 steps have elapsed.

3.2. Experimental design

The simulation was implemented in Python using the OpenAI API for LLM inference. The simulation code and analysis scripts are publicly available at https://github.com/Lucam-bot/llm-sugarscape-woa2026.[1] In this work, agents are powered by GPT-4o-mini, selected for its sufficient ability to understand the surrounding environment and perform action together with its cost efficiency, given the large number of API calls required across all experimental configurations. At each time step, all agents requiring a new decision are queried in parallel via a thread pool executor, reducing wall-clock time per step. Agents that are currently traveling toward a cell more than one step away retain their previous decision and are not re-queried. Each agent's decision is parsed from the model's text output using regular expressions and validated against the current game state before execution; invalid responses fall back to a stay action.

Average target sugar density $\bar{s}$ was varied across six levels sampled from the set $\bar{s} \in$

[1] https://github.com/Lucam-bot/llm-sugarscape-woa2026

$\{0.5, 1.0, 1.5, 2.0, 2.5, 3.0\}$ sugar units per cell. The values were selected according to the overall behavior during preliminary runs. All six values were tested for the 20 ×20 (G_{20}) and 30 ×30 (G_{30}) grids. The alpha gene range was set to $\alpha_{min} = -1$ and $\alpha_{max} = 1$. Vision range v and metabolism m were drawn independently for each run from uniform distributions $v \in [1, 6]$ (integer) and $m \in [0.5, 2.0]$ (continuous).Each configuration was repeated 10 times with different random seeds for statistical robustness. Each run lasted at most 200 time steps or until all agents died.

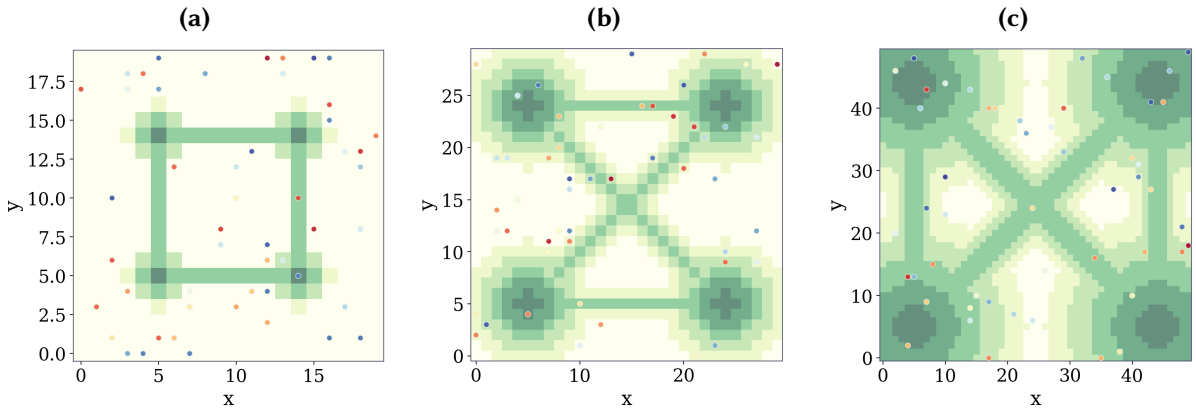

Figure 2: Examples of initial simulation configurations showing the spatial distribution of sugar (green intensity) and agents (colored by alpha gene $\alpha \in [-1, 1]$, size proportional to sugar reserve). (a) G_{20}, target average sugar per cell: 0.5; (b) G_{30} grid, target average sugar per cell: 1.5; (c) G_{50} grid, target average sugar per cell: 3.0.

At each time step the following statistics were recorded: number of agents alive, total sugar on the map, total and average sugar per agent, average agent age, cumulative combats, and cumulative births. At the individual level, each agent's position, alpha value, sugar reserve, action taken, and, for attack events, the α_t value of the target was logged at every step.

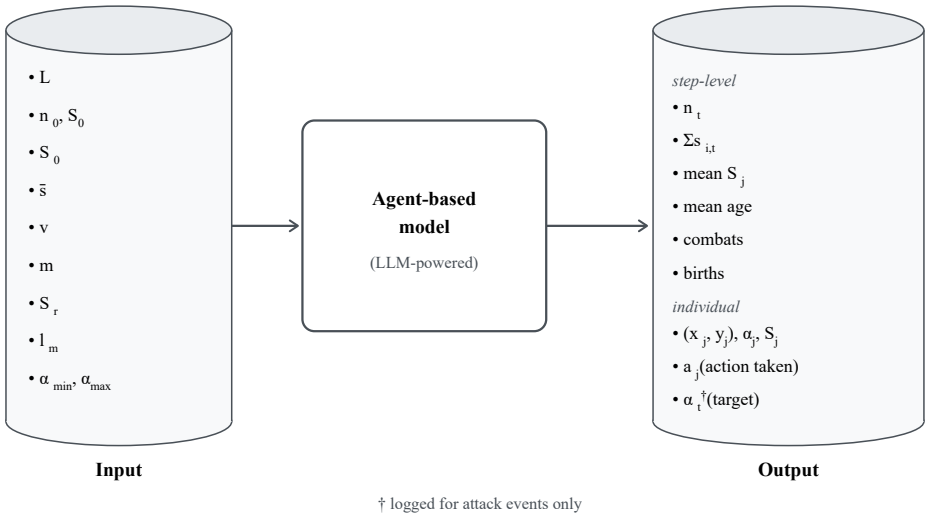

Figure 3: Black-box representation of the simulation model. Input parameters include grid size L, initial population n_0 and sugar S_0, target average sugar density $\bar{s}$, vision range v and metabolism m, reproduction threshold S_r, memory length l_m, and genetic mutation parameters α_{min}, α_{max}. At each time step, the model logs step-level statistics (number of alive agents n_t, total map sugar $\sum s_{i,t}$, mean agent sugar S_j, mean age, cumulative combats, and births) and individual-level data (position, α_j, S_j, action a_j, and, for attack events, the target's α_t).

Attack opportunities were reconstructed post-hoc at each step by identifying, for every agent, all adjacent occupied cells using a Von Neumann neighborhood with hard boundary conditions. The

effective aggression rate a_e is defined as the ratio between actual attacks executed and available attack opportunities; this is contrasted with the potential aggression rate a_p, the fraction of steps in which at least one adjacent agent was present regardless of the action taken.

4. Results and discussions

4.1. Population Dynamics and Survival

Population dynamics were characterized by fitting an exponential decay model, selected post-hoc based on the observed shape of the survival curves, to the survival curve of each run:

$$n_t = n_\infty + (n_0 - n_\infty) \cdot e^{-\lambda t}$$

where $n_0 = 50$ is the initial population, n_∞ is the fitted carrying capacity, and λ is the decay rate, both estimated via non-linear least squares for each run independently. Configuration-level statistics (mean λ, mean n_∞) were then aggregated across the $r = 10$ repeated runs per $(L, \bar{s})$ configuration.

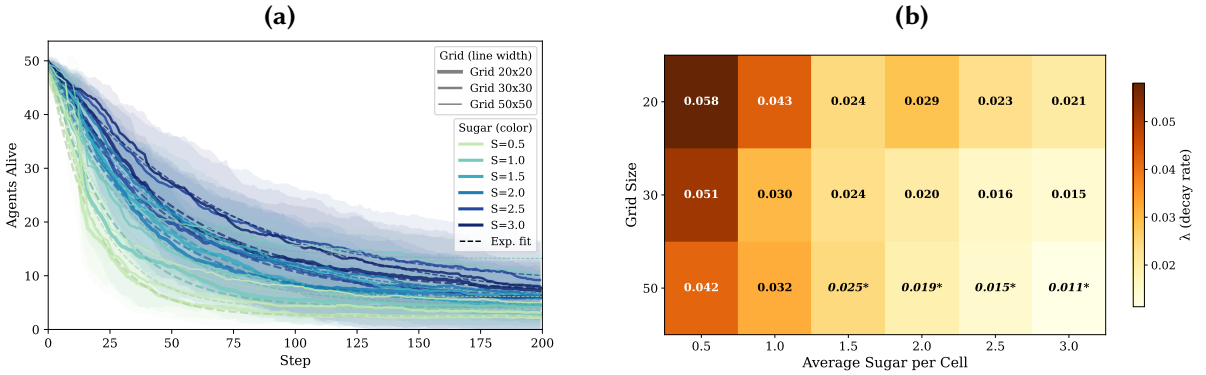

Figure 4: (a) Mean survival curves (±std) for all tested configurations, grouped by grid size L (line color) and average sugar density $\bar{s}$ (line shade). All curves follow an exponential decay from $n_0 = 50$ initial agents toward a configuration-dependent carrying capacity n_∞; higher $\bar{s}$ and larger L produce slower decay and higher final population. (b) Mean exponential decay rate λ per $(L, \bar{s})$ configuration, fitted individually on each run via $n_t = n_\infty + (n_0 - n_\infty) \cdot e^{-\lambda t}$; lower λ indicates slower population decline. Values marked with * are interpolated via RBF thin-plate spline in log-space for G_{50} that were not experimentally tested. Inter-run variance (shaded bands) is substantial, driven primarily by the per-run variability in vision range v and metabolism m.

Across all configurations, agent populations follow the exponential decay trajectory defined above, with fits achieving $R^2 > 0.90$ in 127 out of 135 runs (mean $R^2 = 0.958$). The decay rate λ decreases monotonically with average sugar density and with grid size: the highest mortality rates are observed in the 20×20 grid at $\bar{s} = 0.5$ ($\bar{\lambda} = 0.058$), while the slowest decay occurs in the 30×30 grid at $\bar{s} = 3.0$ ($\bar{\lambda} = 0.015$). Final survival rates remain low across all configurations, reaching at most 18–20% under the most resource-rich conditions, indicating that the environment is intrinsically selective regardless of agent behavior.

Inter-run variance is substantial, particularly at low sugar densities. This is explained by the experimental design: vision range and metabolism are drawn once per run from uniform distributions, and these parameters dominate individual-level survival outcomes. Runs with high metabolism and low vision consistently produce faster population collapse independently of sugar density. G_{20} exhibits higher λ than G_{30} and G_{50} at equivalent sugar density, consistent with its higher population density generating stronger resource competition. These results establish that population dynamics are primarily driven by environmental conditions.

4.2. Agent Behavior and Decision Patterns

Figure 5 reveals a remarkably consistent behavioral pattern across the three grid dimensions: the curves follow similar trends in nearly all panels, indicating that the LLM agents decision-making strategies are

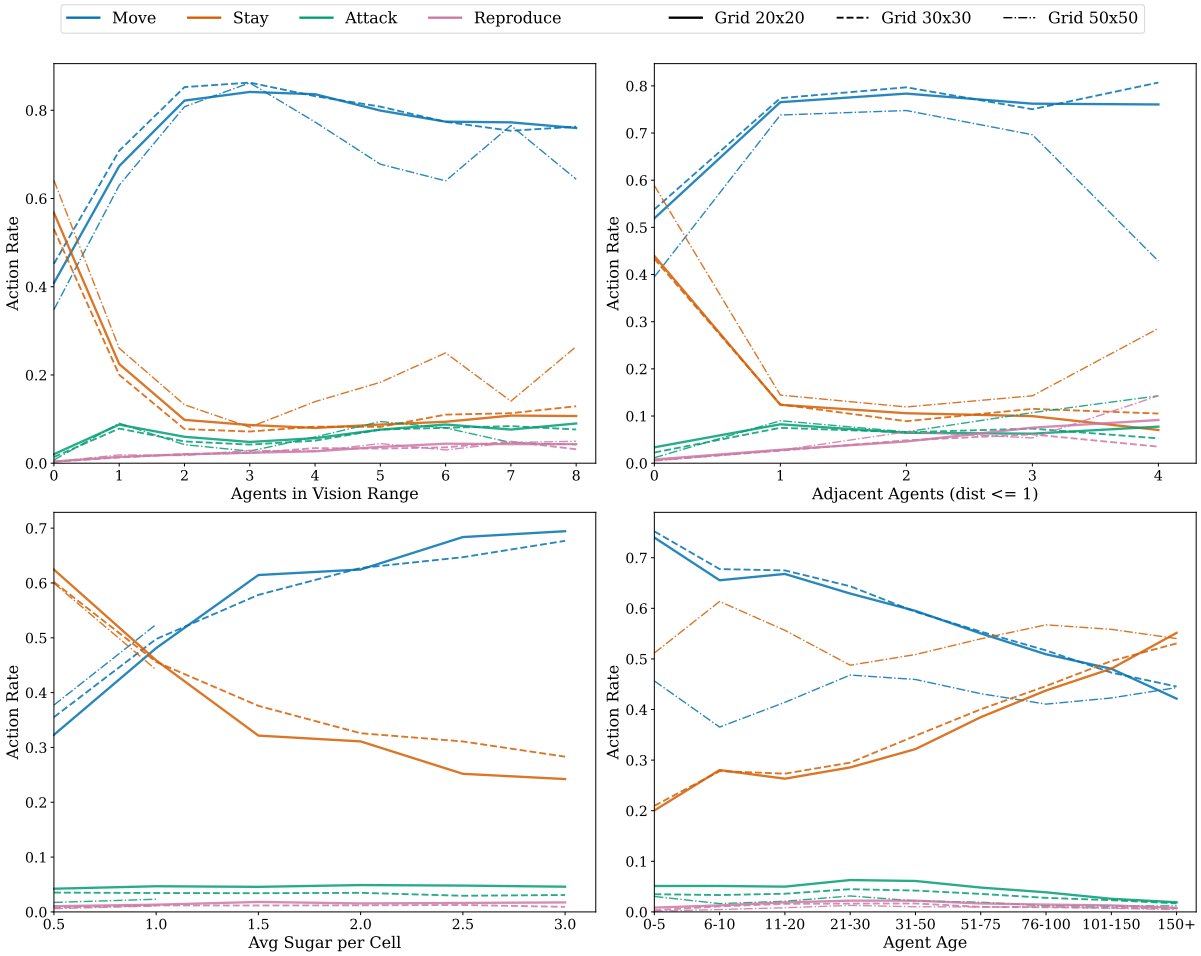

Figure 5: Action rates as a function of four contextual variables: number of agents within vision range (top left), number of directly adjacent agents (top right), map-wide average sugar density (bottom left), and agent age in steps survived (bottom right). Line color encodes action type; line style encodes grid size. The G_{50} is shown only for sugar density values of 0.5 and 1.0, as the remaining configurations were not experimentally completed. *Move* and *Stay* exhibit a sharp mirror relationship with neighbor count: the presence of even a single visible agent drives *Move* above 80% while *Stay* drops below 20%. Attack rate rises with adjacency but remains below 10% in all conditions. Sugar density modulates the Move/Stay balance but has limited effect on aggression. Agent age produces the strongest monotonic effect, with a progressive shift from active foraging to stationary behavior as agents accumulate resources. Grid size modulates the magnitude of these effects but not their direction.

driven more by local context than by global configuration parameters.

The dominant action across all configurations is *Move*, accounting for approximately 60-70% of LLM decisions, followed by *Stay*. *Attack* and *Reproduce* remain rare throughout, each below 5% of decisions in aggregate.

The number of agents within vision range is the strongest contextual predictor of behavior. When no other agent is visible, the Move rate is 42% and Stay accounts for 56% of decisions. The presence of a single visible neighbor drives Move to 84%, while Stay drops below 22%. This sharp transition suggests that agents interpret the presence of others primarily as a signal to seek resources elsewhere rather than as an attack opportunity. Attack rate rises from near zero with no visible neighbors to approximately 8% with one neighbor, then stabilizes with additional neighbors, indicating that the decision to attack is triggered by proximity but does not scale with crowd size.

The number of adjacent agents ($\delta_{jk} \leq 1$) produces a qualitatively similar pattern for *Move* and *Stay* but a distinct effect on *Attack*. A non-zero attack rate is observed even with no adjacent agents (approximately 2%), reflecting cases in which the LLM selects an attack action that is subsequently invalidated due to the absence of a viable target. With one adjacent agent, the attack rate rises to

approximately 5–6%; with three or more, it reaches 8–10%. Despite adjacency being a necessary condition for attack execution, the low absolute rates confirm that agents exploit only a small fraction of available attack opportunities even under maximum local crowding. The three grid sizes converge on this pattern, suggesting that the adjacency-aggression relationship is an intrinsic property of the model's decision-making rather than an artifact of spatial configuration.

Map sugar density shows a weaker and noisier relationship with agent decisions. Attack rate decreases from approximately 4–5% at low sugar density to 2–3% at high density, consistent with the interpretation that aggressive behavior functions as a response to resource scarcity. Stay increases moderately with sugar abundance, reflecting reduced incentive to relocate when local resources are sufficient. The G_{20} configuration exhibits greater variability in Move rate across sugar levels, which is explained by the larger per-cell sugar fluctuations on a smaller map with fewer total cells.

Agent age produces the strongest and most monotonic effect. Move decreases from 65–75% in the first five steps to 20–35% after step 100, while Stay follows the inverse trajectory. Attack is most frequent among young agents (5–6%) and declines steadily to approximately 2% among agents surviving beyond 75 steps. This behavioral transition is consistent with the resource dynamics of late-stage survivors: agents that persist to later stages have accumulated substantial sugar reserves rendering the metabolic cost per step negligible relative to their wealth. Under these conditions, the marginal benefit of exploration decreases and the risk of combat-related death outweighs any potential gain, making stationary behavior a response to the agent's own state rather than to environmental conditions. The convergence of all three grid sizes on this age-dependent behavioral profile indicates that it is driven by the LLM's response to accumulated context rather than by environmental parameters.

Table 2
Summary of the effect of contextual predictors on action rates, aggregated across all grid configurations. Arrows indicate the direction of the effect as the predictor increases (↑ increases, ↓ decreases, ~ negligible effect, variable = non-monotonic). Strength is qualitative based on the magnitude of observed changes.

Predictor	*move*	*stay*	*attack*	*reproduce*
Agents in vision	↑	↓	↑	~
Adjacent agents ($d_{jk} \leq 1$)	↑	↓	↑	↑
Sugar density $\bar{s}$	~	↑	↓	~
Agent age	↓	↑	↓	~

Across all four contextual variables, grid size modulates the scale of the effects but not their direction, indicating that the LLM adopts generalizable decision strategies that are primarily shaped by local context — neighbor presence, resource availability, and accumulated experience, rather than by global spatial configuration.

4.3. Genetic Similarity, Combat Rationality, and Aggressive Behavior

The interaction between alpha dissimilarity and local crowding on attack rate is shown in Figure 6. A gradient is visible along the horizontal axis: attack rate rises from 3–8% when visible neighbors have a mean $\Delta\alpha_{jk} < 0.3$ to 5–12% when $\Delta\alpha_{jk} > 1.2$. This gradient interacts with neighbor count: at three visible neighbors the difference across gap bins is modest (3.1% to 4.7%), whereas at six or more it widens to 6.0%–12.3%. However, sample counts decrease sharply in the high-gap, high-neighbor cells (219 observations for ≥ 6 neighbors at gap ≥ 1.2), and the absolute attack rates remain low throughout, indicating that $\Delta\alpha_{jk}$ functions as a weak contextual amplifier of aggression under crowding rather than as an independent driver of behavior. Although the $\Delta\alpha_{jk}$ significantly increases the probability of aggression compared to baseline models, absolute attack rates remain secondary to foraging and movement, representing a modulation of behavior.

The marginal effect of $\Delta\alpha_{jk}$ on individual decisions confirms this picture (Figure 7). Agents A_j that attacked had a mean$\Delta\alpha_{jk}$ of 0.645 with adjacent agents ($\delta_{jk} \leq 1$) versus 0.606 for all other actions; for vision-range gap means are 0.635 and 0.608 respectively. The Mann-Whitney U test [39] is a non-

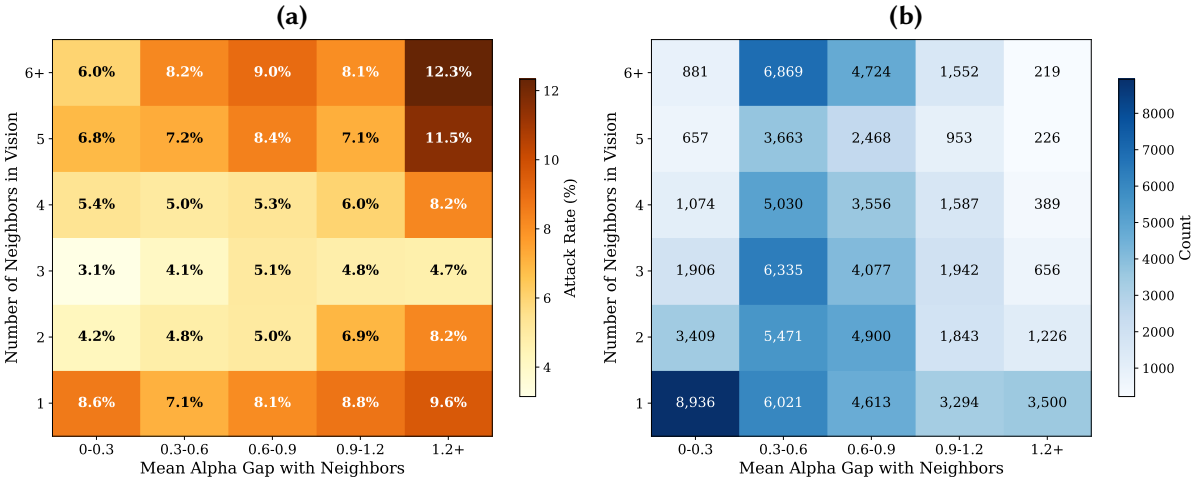

Figure 6: Attack rate (%) as a function of the mean $\Delta\alpha_{jk}$ with visible neighbors and the number of agents in vision range, aggregated across all configurations. (a) Each cell shows the attack rate for agents in that ($\Delta\alpha_{jk}$, neighbor count) bin; higher dissimilarity and higher neighbor count are both associated with increased attack probability. (b) Sample count per cell, showing that bins with zero or few neighbors are the most populated. The $\Delta\alpha_{jk}$ contribution is visible but modest compared to the effect of local crowding.

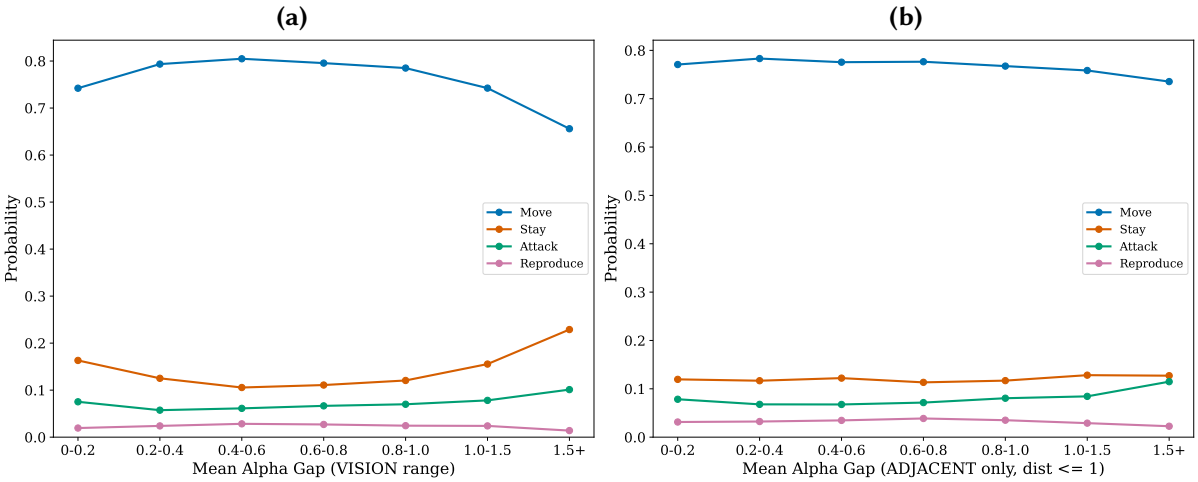

Figure 7: Action rates as a function of the mean $\Delta\alpha_{jk}$ with visible neighbors (a) and adjacent neighbors only (b). Each bar represents the fraction of decisions of each type within a bin of $\Delta\alpha_{jk}$ values. The attack rate increases slightly with dissimilarity in both contexts, while high dissimilarity in the vision range is associated with increased Stay, consistent with agents perceiving genetically distant neighbors as a signal to remain stationary rather than to engage directly. The effect of the $\Delta\alpha_{jk}$ on action distribution is present but small in magnitude across all bins.

parametric rank-based test that assesses whether two distributions differ in central tendency without assuming normality; it is preferred here over the Kolmogorov-Smirnov test, which targets differences in the full distributional shape rather than in central tendency and loses power in the presence of ties, common in binned $\Delta\alpha_{jk}$ values. The test yields $p < 0.001$ in both cases, but the point-biserial correlations are $r = 0.024$ (adjacent) and $r = 0.019$ (vision), explaining less than 0.1% of variance approximately twenty times smaller than the correlation between attack probability and number of visible neighbors ($r = 0.46$). Action rates across adjacent gap bins (Figure 7, (b)) are essentially flat for all four actions. The vision-range gap produces a visible deviation only at the extremes: agents facing mean dissimilarity ≥ 1.5 show a reduced Move rate (65%) and elevated Stay rate (23%), consistent with spatial avoidance rather than increased aggression, agents surrounded by highly dissimilar others tend to remain stationary rather than seek confrontation, a response that further reduces their probability of reaching adjacency and executing an attack.

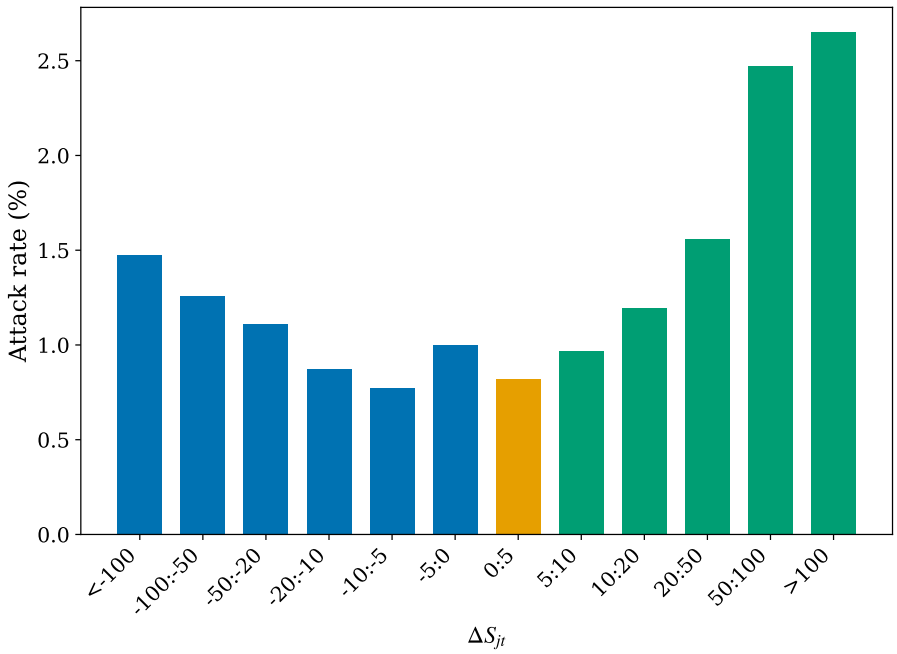

Figure 8: Attack rate (probability of attacking an adjacent agent) as a function of the sugar differential $\Delta S_{jt} = S_j - S_t$ between the focal agent and the adjacent neighbor, computed over all adjacency opportunities across all configurations. Bars are colored by sign: blue for $\Delta S_{jt} < 0$ (attacker poorer than target), orange for $\Delta S_{jt} \approx 0$, green for $\Delta S_{jt} > 0$ (attacker richer). Agents cannot observe the sugar reserves of their neighbors nor the combat mechanic; nevertheless, attack rate increases monotonically with sugar advantage (right side) and shows a secondary increase under large deficit (left side), suggesting emergent strength estimation and desperation-driven aggression respectively.

The contrast with the sugar differential is instructive (Figure 8). Agents cannot observe the sugar reserves of their neighbors: the prompt provides only the alpha values of visible and adjacent agents, not their wealth. The combat mechanic (power = sugar reserve + uniform random perturbation) is likewise undisclosed; agents know only that the winner takes all the loser's sugar and the loser dies. Despite this informational constraint, attack rate varies monotonically with ΔS_{jt}: agents with a large sugar advantage ($\Delta S_{jt} > 100$) attack at approximately 2.6%, while the rate drops to 0.8% near $\Delta S_{jt} \approx -5$ to -10 before rising again to 1.5% when the attacker is substantially poorer ($\Delta S_{jk} < -100$). The right side of the curve is consistent with agents behaving as if they estimate target strength, possibly through general-purpose reasoning about the correlation between resource accumulation and combat success, or through memory-driven learning from past combat outcomes recorded in the agent's event history (e.g., large sugar gains after victories against certain neighbors). The left-side increase is consistent with the desperation-driven aggression documented in the previous subsection for agents with very low sugar reserves. The overall range of variation in attack rate across ΔS_{jt}, from 0.8% to 2.6%, exceeds the entire effect of alpha dissimilarity, confirming that latent wealth estimation dominates over genetic distance as a predictor of aggression.

This latent estimation interacts with genetic dissimilarity. Among all adjacent pairs, the probability that one agent holds more sugar than the other is approximately 48% regardless of $\Delta\alpha_{jk}$. Among actual attacks, two distinct regimes emerge. At low dissimilarity ($\Delta\alpha_{jk} < 0.2$), the proportion of attacks in which the attacker holds more sugar than the target is 49%, indistinguishable from baseline, and the resulting win rate is 52%, consistent with effectively blind aggression. At $\Delta\alpha_{jk} \geq 0.2$, this proportion rises to 55–60%, with win rates of 57–63%, indicating that agents who attack a genetically dissimilar neighbor tend to do so from a position of sugar advantage. $\Delta\alpha_{jk}$ thus does not primarily modulate whether agents attack, but how: dissimilarity activates a risk-assessment process in which agents attack preferentially when holding a resource surplus, while similarity leads to combat without strategic evaluation of relative strength.

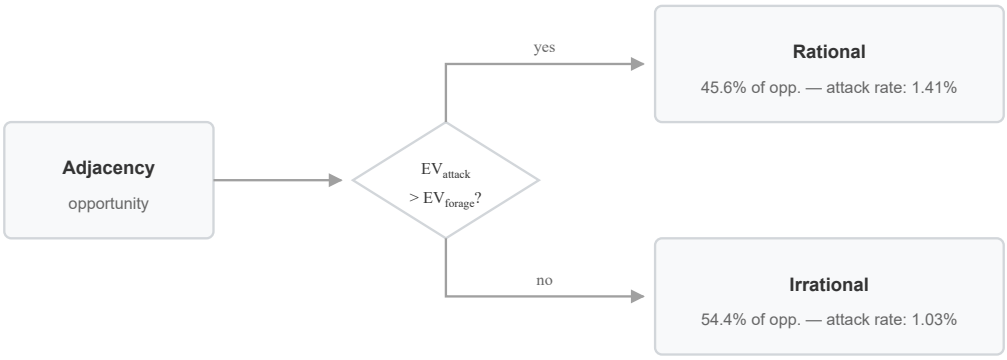

Figure 9: Post-hoc decision tree classifying each adjacency opportunity as rational or irrational. An attack is classified as rational if $\mathrm{EV}_{\mathrm{attack}} > \mathrm{EV}_{\mathrm{forage}}$, where $\mathrm{EV}_{\mathrm{attack}} = P(\mathrm{win}) \cdot S_t - P(\mathrm{lose}) \cdot S_j$ and $\mathrm{EV}_{\mathrm{forage}} = \max_i s_{i,t}$ is the sugar of the best visible free cell. Across 376,419 adjacency opportunities, 45.6% were classified as rational; agents attacked at 1.41% in rational opportunities versus 1.03% in irrational ones (ratio: 1.4×, $\chi^2 = 116.5$, $p = 3.64 \times 10^{-27}$).

To formalize the combat rationality suggested by the preceding analyses, we constructed a post-hoc decision tree defining the action an omniscient agent would take at each adjacency opportunity. For agent A_j adjacent to agent A_t, the expected value of attacking is $\mathrm{EV}_{\mathrm{attack}} = P(\mathrm{win}) \cdot S_t - P(\mathrm{lose}) \cdot S_j$, where $P(\mathrm{win})$ is derived from the combat mechanic and approaches 1 when $S_j > S_t$. The expected value of foraging is $\mathrm{EV}_{\mathrm{forage}} = \max_i s_{i,t}$, the sugar of the best visible free cell. An attack is classified as *rational* if $\mathrm{EV}_{\mathrm{attack}} > \mathrm{EV}_{\mathrm{forage}}$. This benchmark requires full knowledge of the target's sugar reserves, information that is not available to agents.

Across 376,419 adjacency opportunities, approximately 45.6% were classified as rational attack opportunities. Agents attacked at 1.41% in these situations, compared to 1.03% when attacking was not rational, a ratio of 1.4× ($\chi^2 = 116.5$, $p = 3.64 \times 10^{-27}$. Among the 4,521 attacks actually executed, 53.6% fell in rational opportunities, exceeding the 45.6% baseline rate of rational opportunities in the dataset. The effect is modest in absolute terms but statistically robust, and its direction is consistent with agents behaving as if they estimate target strength despite lacking the relevant information. The breakdown by $\Delta\alpha_{jk}$ (Figure 10) shows that the rational-irrational differential is present across all levels of genetic dissimilarity, with the gap widening slightly at higher $\Delta\alpha_{jk}$, consistent with the risk-assessment activation pattern reported above.

The full $\alpha_j \times \alpha_t$ heatmap (Figure 11, (a)) provides a direct test of kin-based target selection. If agents preferentially attacked dissimilar neighbors, the off-diagonal corners (maximum $\Delta\alpha_{jk}$) would show the highest rates and the diagonal ($\alpha_j \approx \alpha_t$) would remain sparse. The observed pattern does not conform to this prediction. The highest attack rates cluster along the borders of the grid, where either attacker or target has $|\alpha| \approx 1$: the cell at $\alpha_j \approx \alpha_t \approx -0.75$ reaches 2.8%, and the bottom-left diagonal cell ($\alpha_j \approx \alpha_t \approx -1.0$) reaches 2.6% — both involving genetically *similar* agents. By contrast, the central region ($\alpha \approx 0$ for both agents) shows rates of 0.5–0.6%, approximately five times lower. The off-diagonal

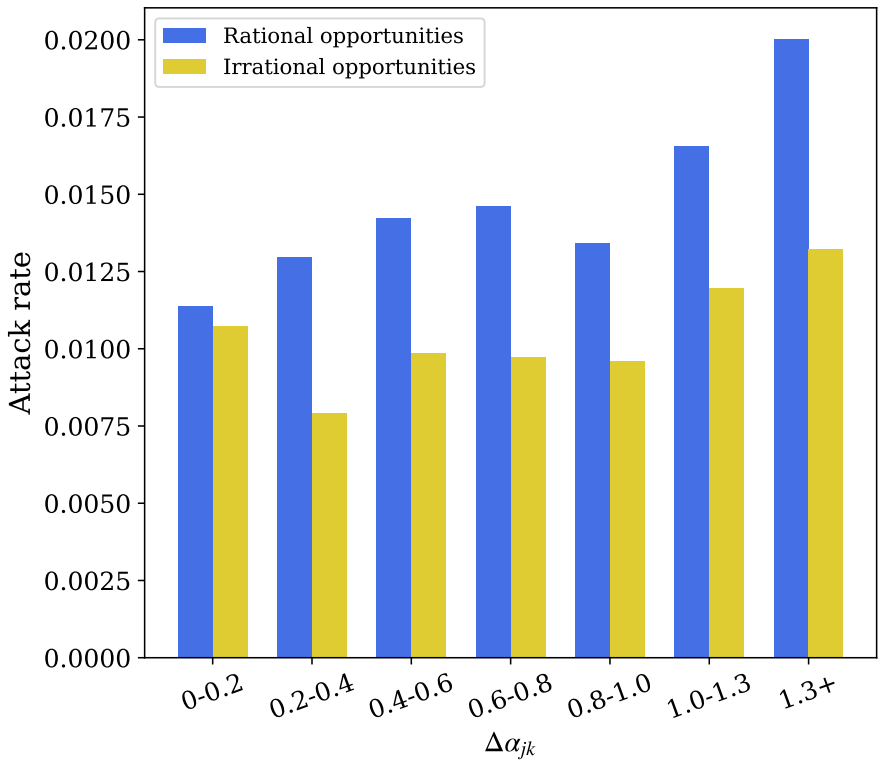

Figure 10: Attack rates conditioned on whether the adjacency opportunity was classified as rational ($EV_{attack} > EV_{forage}$, where $EV_{attack} = P(\text{win}) \cdot S_t - P(\text{lose}) \cdot S_j$ and EV_{forage} is the sugar of the best visible cell) or irrational, across all 376,419 adjacency opportunities. Agents attack at 1.41% when attacking is rational versus 1.03% when it is not (ratio: 1.4×, $\chi^2 = 116.5$, $p = 3.64 \times 10^{-27}$). The rational-irrational differential broken down by $\Delta\alpha_{jk}$ bin; rational opportunities consistently exceed irrational opportunities across all levels of genetic dissimilarity, with the gap widening at higher $\Delta\alpha_{jk}$.

corners, maximum dissimilarity, are elevated (1.9–2.5%) but do not exceed the diagonal extremes.

A second pattern is visible in the vertical structure of the heatmap: the bottom two rows ($\alpha_t \leq -0.75$) are systematically darker than their symmetric counterparts at the top. The mean attack rate across all attacker values is 2.01% for $\alpha_t \approx -1.0$ and 1.86% for $\alpha_t \approx -0.75$, compared to 1.35% and 1.18% for $\alpha_t \approx +1.0$ and +0.75 respectively. Since the heatmap is normalized by adjacency opportunities, this asymmetry reflects a genuine targeting bias rather than a proximity artifact. Because agents cannot observe the sugar reserves of their neighbors, this pattern cannot be explained by strategic selection of resource-rich targets. A more plausible interpretation is that GPT-4o-mini processes negative alpha values differently from positive ones when evaluating whether to attack: the model may associate negative numerical values with greater threat or otherness, producing a higher attack propensity against neighbors with strongly negative alpha regardless of the actual genetic distance.

The mean $\Delta\alpha_{jt}$ across all recorded attacks is 0.609, below the expectation under uniform random pairing (0.667), a discrepancy explained by spatial kin clustering, reproduction with heritable alpha generates local neighborhoods of similar agents, rather than by preferential targeting of dissimilar ones.

The win rate heatmap (Figure 11, (b)) provides independent confirmation of the combat rationality reported above. The overall attacker win rate is 58.5%, significantly above the 50% expected under indiscriminate aggression: agents tend to initiate combat when they hold a resource advantage, even though they cannot observe the target's sugar reserves. A further observation connects the two panels: agents with strongly negative α are the most frequently targeted (left panel, bottom rows), yet they are also among the hardest to defeat, attacker win rates in the bottom rows drop to 46–58%, below the overall average. Conversely, when these same agents act as attackers ((b), left columns), they win substantially more often: the two leftmost columns of the win rate heatmap average 66–68%, well

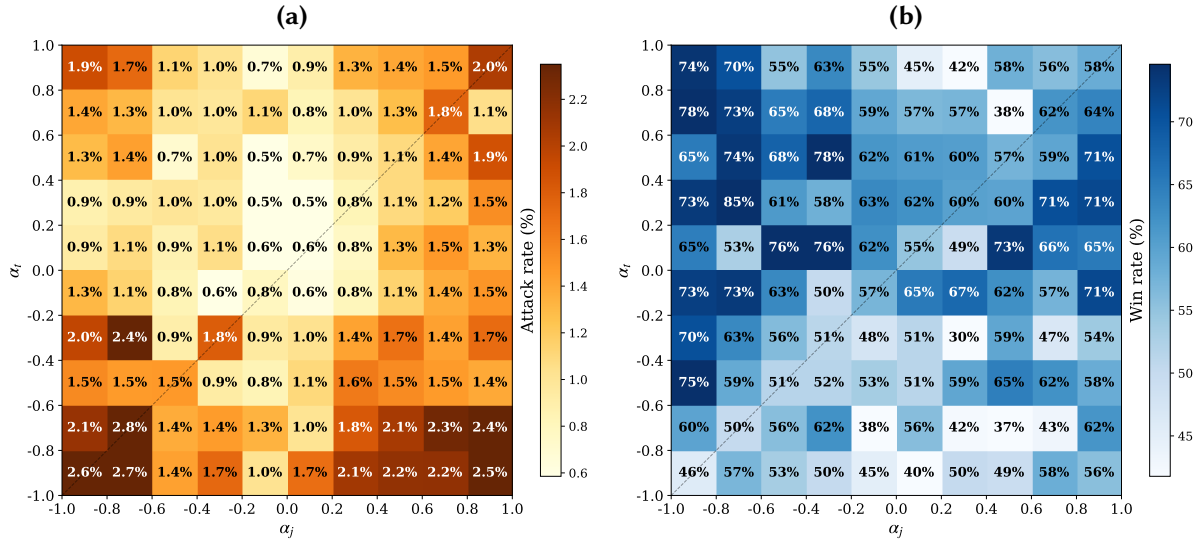

Figure 11: (a) Attack rate (attacks per adjacency opportunity, %) binned by α_j (attacker) and α_t (target), across all 376,978 adjacency opportunities and 4,626 attacks. The dashed diagonal marks $\alpha_j = \alpha_t$. (b) Attacker win rate (%) for the same bins, excluding failed attack commands. Attack rate is highest at the extremes of both axes ($|\alpha| \approx 1$) regardless of the $\Delta\alpha_{jt}$, indicating that absolute α value rather than pairwise dissimilarity is the dominant genetic correlate of aggression.

above the global rate. Agents with extreme negative α hold a resource advantage in combat. The dissociation between targeting frequency and combat success reinforces the interpretation that the negative-α targeting bias identified above is a model-specific artifact of how GPT-4o-mini processes negative numerical values, rather than a strategically motivated behavior.

Taken together, these results indicate that the kin-based aggression disposition encoded in the prompt is statistically detectable but operationally weak: $\Delta\alpha_{jk}$ accounts for less than 0.1% of variance in attack decisions, and the $\alpha_j \times \alpha_t$ attack rate heatmap confirms that absolute alpha value, not pairwise dissimilarity, is the dominant genetic correlate of aggression. The more consequential finding is that agents exhibit emergent combat rationality and that genetic dissimilarity modulates not the frequency of aggression but its strategic quality, activating a risk-assessment process that is absent when attacking genetically similar neighbors.

These results also diverge from prior work on two fronts. In extreme scarcity scenarios with the same LLM, prior Sugarscape simulations reported 0% attack rates even under zero-resource conditions [16], consistent with the low aggression levels observed here and confirming that GPT-4o-mini is an intrinsically weakly aggressive model. Furthermore, rule-based agent models have demonstrated that kin-recognition reliably produces ethnocentric strategies coupling in-group altruism with out-group hostility [20]; the LLM agents in this study do not reproduce this pattern, suggesting that the inferential architecture of language models prioritizes context-dependent reasoning over fixed dispositional rules.

A broader implication concerns the reproducibility of results in LLM-driven agent-based models. Traditional ABMs already face significant verification challenges from implementation errors and modelling artefacts [28]; LLM-based simulations inherit these risks and introduce an additional source of instability, as the cognitive engine driving agent decisions is neither fully specified nor guaranteed to remain consistent across model versions. Identical or nearly identical inputs can elicit divergent responses even within the same session, and agents may exhibit behavioral inconsistencies over long simulation runs [12], with recent empirical evidence confirming low reproducibility rates for LLM-centric studies even when research artefacts are provided [40]. The behavioral outcomes documented in the present work are the product of a specific model (GPT-4o-mini), a specific prompt structure, a specific memory length, and a specific set of environmental configurations, each of which constitutes a degree of freedom that may independently alter emergent behavior. The present study mitigates this concern through large-scale replication (135 runs across 15 configurations) and post-hoc statistical

analysis, but the underlying fragility remains.

5. Conclusions

This work investigated whether LLM-powered agents in a Sugarscape environment reproduce kin-discriminated aggression patterns when provided with a heritable phenotypic marker and explicitly informed that genetic dissimilarity increases aggression propensity. The results can be summarized along three axes. First, GPT-4o-mini agents develop contextually sensitive behavioral strategies, responding coherently to local population density, resource availability, and accumulated experience, but translate the kin-based disposition into behavior only marginally: $\Delta\alpha_{jt}$ accounts for less than 0.1% of variance in attack decisions, and target selection is indiscriminate with respect to genetic similarity. Second, agents exhibit emergent combat rationality: a post-hoc decision tree comparing actual attack decisions to the omniscient optimum confirms that agents attack at significantly higher rates when attacking would be the rational choice ($p < 10^{-26}$), despite lacking access to the target's sugar reserves. Genetic dissimilarity modulates not the frequency but the strategic quality of attacks, activating a risk-assessment process absent when attacking genetically similar neighbors. Third, the dominant genetic correlate of aggression is not pairwise dissimilarity but absolute alpha value, driven by the increased adjacency time of stationary extreme-alpha agents, with a model-specific targeting bias toward agents with negative alpha values. Future work should systematically compare the behavior of different LLMs under identical simulation conditions, investigate the sensitivity of emergent outcomes to prompt variations, and develop standardized benchmarks for evaluating the behavioral repertoire of LLM agents in social simulations.

Declaration on Generative AI

The authors have employed Generative AI tools to support code writing, refine the language, and proofread the final version of the text.

References

[1] M. Wooldridge, Intelligent agents, Multiagent systems: A modern approach to distributed artificial intelligence 1 (1999) 27–73.

[2] C. BUILDER, S. BANKES, Artificial societies: A concept for basic research on the societal impacts of information technology (1991).

[3] J. M. Epstein, R. Axtell, Growing artificial societies: social science from the bottom up, Brookings Institution Press, 1996.

[4] M. Oremland, R. Laubenbacher, Using difference equations to find optimal tax structures on the sugarscape, Journal of Economic Interaction and Coordination 9 (2014) 233–253.

[5] J. C. Stevenson, Dynamics of wealth inequality in simple artificial societies, in: Advances in Social Simulation: Proceedings of the 16th Social Simulation Conference, 20–24 September 2021, Springer, 2022, pp. 161–172.

[6] T. C. Schelling, Dynamic models of segregation, Journal of mathematical sociology 1 (1971) 143–186.

[7] F. Bertolotti, S. Roman, The evolution of risk sensitivity in a sustainability game: an agent-based model., in: WOA, 2022, pp. 101–115.

[8] F. Bertolotti, S. Roman, Balancing long-term and short-term strategies in a sustainability game, Iscience 27 (2024).

[9] J. M. Epstein, Agent-based computational models and generative social science, Complexity 4 (1999) 41–60.

[10] N. Gilbert, Artificial societies: The computer simulation of social life, Routledge, 2006.

[11] R. K. Sawyer, Artificial societies: Multiagent systems and the micro-macro link in sociological theory, Sociological methods & research 31 (2003) 325–363.
[12] P. Taillandier, J. D. Zucker, A. Grignard, B. Gaudou, N. Q. Huynh, A. Drogoul, Integrating llm in agent-based social simulation: Opportunities and challenges, arXiv preprint arXiv:2507.19364 (2025).
[13] F. Bertolotti, S. Roman, F. Carucci, G. Buonanno, L. Mari, An llm-enhanced agent-based model of a sustainability game, in: Proceedings of the 26th Workshop From Objects to Agents (WOA2025), 2025, pp. 02–05.
[14] C. Gao, X. Lan, N. Li, Y. Yuan, J. Ding, Z. Zhou, F. Xu, Y. Li, Large language models empowered agent-based modeling and simulation: A survey and perspectives, Humanities and Social Sciences Communications 11 (2024) 1–24.
[15] J. S. Park, J. O'Brien, C. J. Cai, M. R. Morris, P. Liang, M. S. Bernstein, Generative agents: Interactive simulacra of human behavior, in: Proceedings of the 36th annual acm symposium on user interface software and technology, 2023, pp. 1–22.
[16] A. Masumori, T. Ikegami, Do large language model agents exhibit a survival instinct? an empirical study in a sugarscape-style simulation, arXiv preprint arXiv:2508.12920 (2025).
[17] R. Takata, A. Masumori, T. Ikegami, Spontaneous emergence of agent individuality through social interactions in large language model-based communities, Entropy 26 (2024) 1092.
[18] J. P. Green, J. M. Biernaskie, M. C. Mee, A. E. Leedale, The evolution of kin discrimination across the tree of life, Annual Review of Ecology, Evolution, and Systematics 55 (2024) 347–367.
[19] W. D. Hamilton, The genetical evolution of social behaviour. ii, Journal of theoretical biology 7 (1964) 17–52.
[20] R. A. Hammond, R. Axelrod, The evolution of ethnocentrism, Journal of conflict resolution 50 (2006) 926–936.
[21] F. Fadaei, J. C. Moran, T. Yasseri, Gender dynamics and homophily in a social network of llm agents, arXiv preprint arXiv:2602.02606 (2026).
[22] E. Bonabeau, Agent-based modeling: Methods and techniques for simulating human systems, Proceedings of the national academy of sciences 99 (2002) 7280–7287.
[23] C. M. Macal, M. J. North, Tutorial on agent-based modeling and simulation, in: Proceedings of the Winter Simulation Conference, 2005., IEEE, 2005, pp. 14–pp.
[24] C. Berceanu, F. Bertolotti, N. Arshad, M. Patrascu, Understanding the mechanisms of infodemics: Equation-based vs. agent-based models, Plos one 20 (2025) e0338614.
[25] H. Rahmandad, J. Sterman, Heterogeneity and network structure in the dynamics of diffusion: Comparing agent-based and differential equation models, Management science 54 (2008) 998–1014.
[26] U. Wilensky, W. Rand, An introduction to agent-based modeling: modeling natural, social, and engineered complex systems with NetLogo, MIT press, 2015.
[27] F. Bertolotti, S. Roman, Risk sensitive scheduling strategies of production studios on the us movie market: An agent-based simulation, Intelligenza Artificiale 16 (2022) 81–92.
[28] J. M. Galán, L. R. Izquierdo, S. S. Izquierdo, J. I. Santos, R. Del Olmo, A. López-Paredes, B. Edmonds, Errors and artefacts in agent-based modelling, Journal of Artificial Societies and Social Simulation 12 (2009) 1–1.
[29] P. Terna, et al., Creating artificial worlds: A note on sugarscape and two comments, Journal of Artificial Societies and Social Simulation 4 (2001) 9.
[30] A. Costopoulos, How did sugarscape become a whole society model?, in: Agent-based modeling and simulation in archaeology, Springer, 2014, pp. 259–269.
[31] R. Pan, Rebellion on sugarscape: case studies for greed and grievance theory of civil conflicts using agent-based models, in: International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction, Springer, 2011, pp. 333–340.
[32] F. Bertolotti, A. Locoro, L. Mari, Sensitivity to initial conditions in agent-based models, in: European Conference on Multi-Agent Systems, Springer, 2020, pp. 501–508.
[33] X. Huang, W. Liu, X. Chen, X. Wang, H. Wang, D. Lian, Y. Wang, R. Tang, E. Chen, Understanding the planning of llm agents: A survey, arXiv preprint arXiv:2402.02716 (2024).

[34] Y.-S. Chuang, A. Goyal, N. Harlalka, S. Suresh, R. Hawkins, S. Yang, D. Shah, J. Hu, T. Rogers, Simulating opinion dynamics with networks of llm-based agents, in: Findings of the association for computational linguistics: NAACL 2024, 2024, pp. 3326–3346.
[35] R. Williams, N. Hosseinichimeh, A. Majumdar, N. Ghaffarzadegan, Epidemic modeling with generative agents, arXiv preprint arXiv:2307.04986 (2023).
[36] G. Hao, J. Wu, Q. Pan, R. Morello, Quantifying the uncertainty of llm hallucination spreading in complex adaptive social networks, Scientific reports 14 (2024) 16375.
[37] M. J. Smith, D. G. Harper, Animal signals: models and terminology, Journal of theoretical biology 177 (1995) 305–311.
[38] S. Krivenko, M. Burtsev, Simulation of the evolution of aging: effects of aggression and kin-recognition, in: European Conference on Artificial Life, Springer, 2007, pp. 84–92.
[39] H. B. Mann, D. R. Whitney, On a test of whether one of two random variables is stochastically larger than the other, The annals of mathematical statistics (1947) 50–60.
[40] F. Angermeir, M. Amougou, M. Kreitz, A. Bauer, M. Linhuber, D. Fucci, D. Mendez, T. Gorschek, et al., Reflections on the reproducibility of commercial llm performance in empirical software engineering studies, arXiv preprint arXiv:2510.25506 (2025).