https://doi.org/10.1038/s44260-026-00075-1

Too human to model: the uncanny valley of large language models in simulating human systems

Yongchao Zeng[1] ✉, Calum Brown[1] & Mark Rounsevell[1,2,3]

Large language models (LLMs) have been increasingly used to simulate human behaviour because of their ability to generate contextually coherent dialogues. Such abilities can enhance the realism of models. However, the pursuit of realism is not necessarily compatible with the epistemic foundation of modelling. We explore when LLM agents can be too 'human' to model, i.e., when they are too expressive, detailed and intractable to be consistent with the abstraction, simplification, and interpretability typically demanded by modelling. Through a model-building thought experiment, we uncover five core dilemmas: a temporal resolution mismatch between natural conversation and abstract time steps; the need for intervention in agent conversations without undermining spontaneous outputs; the temptation to introduce rules while maintaining conversational naturalness; the tension between role consistency and role evolution; and the challenge of understanding emergence. These dilemmas lead LLM agents to an "uncanny valley": more realistic than rule-based agents but recognisably unhuman.

Large language models (LLMs), such as GPT[1] and Claude[2], have been increasingly used to simulate human systems[3–5]. With the capability of generating coherent, human-like text, LLMs enable researchers to build and explore artificial societies comprising a variety of autonomous agents[4]. LLM-based agents can express intent, respond to requests, and even mimic complex social behaviours, which are often implemented through well-designed prompts and integration of multiple agents[6]. This fosters the emergence of narrative-rich simulation[7–9]–a system that can operate through natural language, rather than relying on predefined state machines or rule-based logic.

The development of LLMs is exciting. LLMs seemingly provide new features that conventional, rule-based agents often struggle with: plausible, situated, contextually rich behaviour. Literature shows that, through LLMs, simulation can capture the texture of human life[7], ranging from daily conversations and social norms[10] to improvisation and persuasion[11]. Numerous recent papers demonstrate that in LLM-driven social environments, agents can negotiate dinner plans, discuss local events, or co-compose storylines (see ref. 4 for a comprehensive review).

However, LLMs have many widely known technical limitations, such as hallucination[12–14], bias[15–18], and security issues[19,20], which may affect their integration within social simulation. Beyond these, we argue that there is a more profound methodological and epistemological tension underlying LLM-driven simulation, even if all these technical limitations are solved: LLM agents can speak and act like real humans, but this realism comes in social simulation by risking model validity, interpretability, and explanatory rigour. The core idea that is explored in this paper is that "LLM agents are often too human to model." Their simulation of human behaviours is too detailed to serve as an abstract representation, but too artificial to serve as an explanation. Their use, therefore, conflicts with the very nature of social modelling as an epistemic tool for extracting insights from the noisy real world[21–26].

This tension results from a fundamental mismatch between how models abstract social processes and how LLMs generate expressive behaviour. Where conventional agents evolve following clearly defined rules[27], LLM agents generate narratives conditioned on textual inputs[7]; where conventional simulation aims to distil mechanisms[28], LLM agents necessitate sophisticated external mechanisms designed to enable emotional change, memory management, and reasoning processes[29]. Furthermore, while rule-based agents operate on abstract temporal frameworks (e.g., interacting yearly[30]), LLM agents respond through high-resolution, turn-by-turn dialogues[4]. Such discrepancies make the causal linkages between microscopic agent behaviours and system-level social emergence unclear.

The purpose of this paper is to reflect critically on the trend of embedding LLM agents in social simulation without sufficient

[1]Institute of Meteorology and Climate Research, Atmospheric Environmental Research (IMK-IFU), Karlsruhe Institute of Technology, Garmisch-Partenkirchen, Germany. [2]Institute of Geography and Geo-ecology, Karlsruhe Institute of Technology, Karlsruhe, Germany. [3]School of Geosciences, University of Edinburgh, Drummond Street, EH25 9RG Edinburgh, UK. ✉e-mail: yongchao.zeng@kit.edu

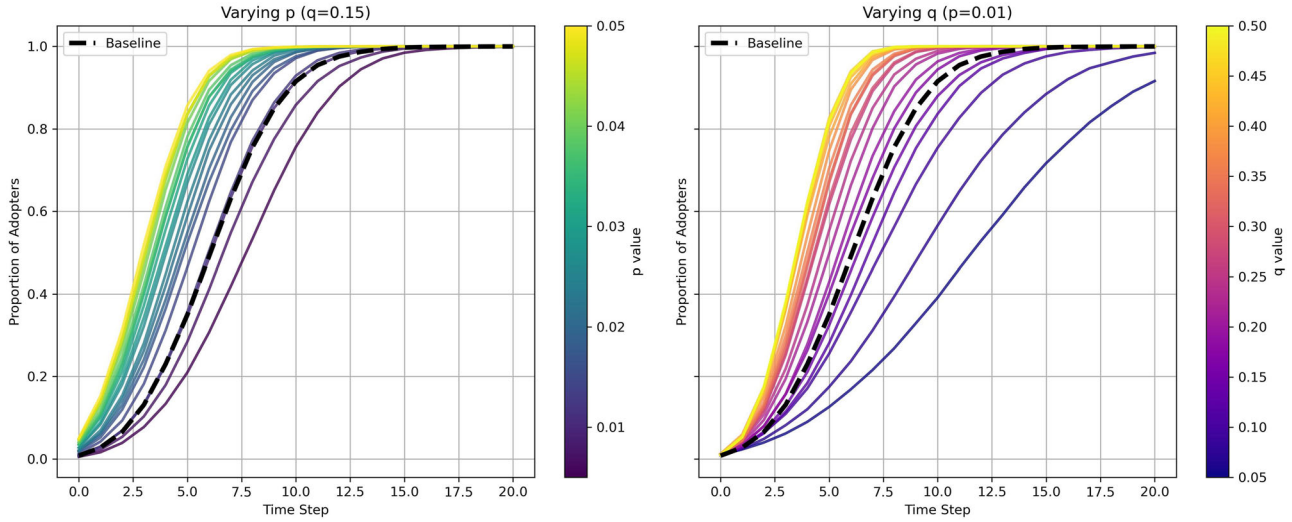

Fig. 1 | The parameters of the baseline curve are p = 0.01 and q = 0.15. Manipulating the values of p and q alters the shape of the diffusion curve: A higher p accelerates early adoption, making the inflection point happen earlier. A higher q amplifies adoption later in the process, steepening the curve after the inflection point.

consideration of the methodological legitimacy. We use a classic innovation diffusion model–the Bass model[31] – to conduct a thought experiment demonstrating the consequences of the epistemic mismatch. Through the process of building an LLM-driven version of the Bass model, we elaborate on how LLM agents' realism gradually introduces liabilities, and how reducing such liabilities may result in an "uncanny valley of agents"–not abstract enough to clarify underlying social mechanisms, while not realistic enough to naturally exhibit human behaviours.

We do not advocate abandoning LLMs in simulating human systems, but propose applying generative AI agents to domains where they fit better. Based on the thought experiment, we make a preliminary attempt to position LLMs using the framework of different modelling purposes[21] and tease out the "comfort zones" of LLMs to spur attention to the evolution of LLM technologies as well as how they reshape our understanding of their roles in simulation science.

A brief revisit of the classic Bass diffusion model and its epistemic strength

Since Frank Bass first introduced his diffusion model in 1969[31], it has been one of the most enduring, widely used approaches to modelling innovation diffusion processes. From predicting the popularisation of colour television[31] to modelling the adoption of renewable energy technology[32], the Bass diffusion model provides a simple but powerful tool for researching how innovation, such as new thoughts, behaviours, technologies, and products, spread in society[33]. The model's power does not lie in its sophistication but in its simplicity and elegance[34]: using only two core parameters to capture various diffusion dynamics in the real world. This model is certainly not flawless, but its structural clarity and interpretability make it a good illustrative example of how to build models. Before proceeding to build an LLM-based diffusion model, we briefly revisit the classic model here.

In the model's differential equation form, the proportion of new adopters $f(t)$ is calculated as:

$$f(t) = \left[p + q \cdot F(t)\right] \cdot \left[1 - F(t)\right] \quad (1)$$

where $F(t)$ is the cumulative proportion of adopters up to time t; p and q respectively represent:

1. External influence that causes individuals to adopt an innovation, such as the influence of advertisements and mass media.

2. Internal influence that leads individuals to adopt, such as the word-of-mouth effect, imitation, or peer pressure.

The intuition is straightforward: an individual adopts an innovation due to the influence of social media and the people connecting with that individual. Based on this intuition, the model can be naturally transformed into an agent-based model. Here is a one-sentence description of the core micro mechanism of the agent-based diffusion model:

In a social network, an agent can become an innovation adopter following the probability p (influence of mass media) and probability q (influence of network neighbours) in every simulation step, which often represents one year.

This can be expressed as a simple IF-THEN rule that drives an agent to become an adopter[35]:

$$\text{IF} x_1 < p \; or \; x_2 < q\frac{n_{i,a}}{n_i} \text{ THEN agent } i \text{ becomes an adopter} \quad (2)$$

where x_1 and x_2 are numbers randomly sampled from a uniform distribution between 0 and 1; n_i is the number of neighbours of agent i; $n_{i,a}$ is the number of adopters among the agent's neighbours.

The agent-based version not only preserves the clarity and interpretability of the original model but also offers higher flexibility to consider spatial heterogeneity[30,36,37], network structure[38,39], individual attributes[40,41], and decision theories/algorithms[42–45]. It helps researchers to gain insights into agent behaviour and emergent diffusion curves[46]. Figure 1 displays how the shape of the diffusion curve changes under different p and q values.

Constructing an LLM-driven diffusion model: A modelling thought experiment

To vividly reveal the tension between LLM-driven social simulation and the abstraction principle of simulation, we carry out a hypothetical modelling exercise: building an LLM version of the Bass diffusion model, with the purpose of understanding how individual adoption decisions lead to the emergence of the diffusion patterns. We do not intend to develop a fully functional simulation system, nor to exhaustively explore all possible modelling choices. Instead, we conduct a modelling thought experiment and highlight critical modelling decisions that introduce a new methodological dilemma. Through this process, we present how LLMs' advantages (especially their expressiveness and contextually sensitive behaviour) turn out to be liabilities undermining the utility of modelling and simulation as an epistemic tool.

A promising start–defining agent types

We begin with a familiar setup. Drawing from Rogers' typology of innovation adopters[33], we create an artificial society incorporating 100 consumer agents, which can be divided into five typical groups: innovator, early adopter, early majority, late majority, and laggards[33]. Each group has distinct tendencies toward risk, conformity, and social influence. To give these agents identities, we can compose personalised prompts for each individual agent. For instance, an innovator named "Novy" is a tech enthusiast who is keen to be the first to try new technologies, while a laggard named "Stony" is a cautious, habit-bounded traditionalist (here we try to avoid too specific a description to mitigate stereotyping, which is another crucial issue but not within the discourse here). Such prompt design does not intend to merely label the agents but to inject personality traits that shape their conversation and decision-making. In addition to the consumer agents, we can incorporate an advertiser agent that generates advertisements accessible by other agents. A simple alternative way is to predefine a list of text representing advertisements and broadcast the text periodically to inform the consumer agents during simulations. Either way, the purpose is to schedule spreading advertisements that may increase consumer awareness and perception of a certain new product.

At this stage, the process seems intuitive. Instead of assigning cold parametric numbers, we are creating "animated characters" with decisions that derive from contextually rich interaction. In principle, this should allow us to explore *how* adoption emerges in far more detail than with a simple numerical model or abstract rule-based agents–an apparently promising avenue towards high realism.

Dilemma I: "Small feet in big shoes"—the temporal resolution mismatch

A complication appears immediately when we proceed to decide the time scale of agent interactions. In conventional diffusion models, time is abstracted into discretised steps that often indicate months or years, during which agents can interact and change their states. This temporal compression is fundamental to the tractability and interpretability of simulation: one step corresponds to a macroscopic social effect, e.g., the change in adoption rate.

However, LLM agents operate on a radically different basis in terms of temporal resolution. Natural language is highly fine-grained and situated. Normally, conversations unfold in seconds or minutes rather than years. Hence, when we arrange a yearly conversation between, e.g., "Novy" and "Stony", we are in practice matching microscopic social behaviours with a macroscopic temporal framework. The result is awkward misalignment and leads to the first modelling dilemma: Can Novy and Stony's interesting dialogue truly represent a whole year's change in thoughts? If not, should we refactor the simulation to cater to the temporal resolution of real conversations? Neither choice is satisfactory: the former demands questionable cognitive compression, while the latter may cause unmanageable computational costs and massive textual outputs, most of which are trivial.

In contrast to simulating consumer interactions via fine-grained conversations, it is more likely that modellers adopt a technique or workaround to "catalyse" yearly change in consumer thoughts through limited times of naturally unfolded agent conversations, i.e., to compress the cognition. To avoid our model-building process being stuck at this step, let us assume such cognitive comprehension is favoured due to relatively higher feasibility, and a one-time conversation can represent a year of conversations between agents. This assumption does not mean cognitive compression is easy to implement in practice. Instead, it should be noted that there is no established method or validation criteria regarding cognitive compression using LLMs, and the assumed feasibility is simply to let us move forward with the thought experiment, which, nevertheless, leads directly to the next dilemma.

Dilemma II: "No time for small talk!"—too naturalistic to avoid intervention

How do agents decide to adopt the innovation, and what interactions can lead to the decision? In the Bass diffusion model, innovation adoption is influenced by two key parameters–the innovation coefficient and the imitation coefficient, which together determine the rate of adoption in each step[31]. This process is simplified, abstract, but transparent. However, for Novy and Stony, how do they decide whether to adopt the innovation through a yearly conversation? How many rounds of conversation are they allowed to have in order to naturally reach the yearly decision? One? Two? Ten? Or let the agent generate a cue, such as "1" representing adoption and "0" no adoption, without constraining the number of conversation rounds? Allowing the agents to converse without artificially imposed limits on conversations seems to be a good choice, as it gives room for a decision to emerge. However, this approach does not guarantee that these stochastic agents will not be trapped in a long or even endless conversation loop. Neither does it mean that the eventual timing and cause of adoption will represent those of the real-world situation being modelled. Therefore, a compromise is to let the agents make a clear decision within a certain number of rounds of conversation; otherwise, their decisions remain unchanged. We can prompt them to do so, add a programme to cut their conversation, create another functional agent, e.g., a conversation organiser, to monitor their decisions, or apply all of these approaches. However, we have to deliberately intervene in their conversation to avoid unexpected issues.

Since Novy and Stony can only speak once "per year" within limited turns of conversation, we must let them be clear about what their conversation is focused on. That is, to encourage or discourage them from communicating some content. In real-world conversations between acquaintances, it is normal to talk about everyday life topics, such as the weather, gossip, and frustration, within which innovation adoption is embedded. Indeed, LLM agents can generate such socially realistic conversations with ease – they actually excel in mimicking fine-grained human communications[47]. However, from the perspective of modelling, such content is noise: LLMs cannot embed meaningful cues in such exchanges, and so they become irrelevant, and if agents spend most of their interactions on irrelevant chatter, the model loses its interpretability and efficiency. That means that we should prompt agents to focus their interactions on the new technology, for example, "Your goal is to evaluate the new product and decide whether to adopt it through 10 rounds of conversation." The risk of doing so is that we may script the simulation too tightly, suppressing the spontaneous behaviour of LLM agents we expected to capture initially. It is noteworthy that the new product itself should also be described in detail to contextualise conversations that are naturally unfolded. Furthermore, the form, content, and timing of advertisements should not be abstract nor ignored, as they are treated in the original Bass diffusion model. The requirement for realism seems to be forceful rather than selectable if modelling is centred around fine-grained conversations, while the necessity of simplification and/or abstraction appears unavoidable.

We are now faced with a tricky dilemma: the boundary between social realism and epistemic noise. If the naturalness of the LLM agents' interaction is insufficient, their advantage over rule-based agents becomes dubious; if the naturalness is excessive, the core decision logic of LLM agents is buried in linguistic redundancy. If we accept insufficient naturalness, we are forced to intervene: shortening conversations, filtering content, and constraining focus. All of these interventions make the simulation move from "modelling decisions" towards "staging decisions". Nevertheless, if we allow for excessive naturalness, feasibility issues arise even earlier than methodological challenges. A balanced naturalness is even more difficult to justify, as "balanced" itself is obscure and often context-dependent.

Again, to move forward, let us choose the former approach for higher feasibility.

Dilemma III: "I respect rules, but don't 'rule' me" —the temptation of resorting to rules

Now we have 100 LLM agents with well-defined profiles, able to communicate yearly with a focused topic: adopting a new technology. Meanwhile,

their conversations are strictly limited within a certain length to avoid inefficiency in outputs. It is time to design how these agents make final (yearly) decisions based on historical conversations.

It is certainly viable to prompt the LLM agents to make decisions based on a set of criteria, such as considering the number of friends that have adopted the innovation and the price against a threshold, or the times a consumer agent has been exposed to advertisements. However, this approach suggests that the conversations between these agents are unnecessary–we only need to inform the LLM agents about some numerical indicators, which means the simulation is reduced to a rule-based model, and all the modelling decisions we made about conversation mechanisms are irrelevant. In practice, this approach is not only feasible but also intuitive and straightforward, which does not require the agents to converse, memorise, or even think. However, we are using LLMs in this example in order to model the emergence of such outcomes, not to prescribe them. To mitigate the risk of reverting to abstract rule-based modelling, we have to let the LLM agents' decisions form naturally from historical conversations and probably advertisements received, rather than relying on a set of IF-THEN rules. This leads to the next problem: LLMs do not have persistent memory.

Without intervention, it is very likely that two LLM agents' conversation in a certain year takes no account of anything they have discussed in previous years. This lack of hindsight results in repetition across conversations. To endow the agents' behaviour with continuity, we have to build a memory system with all received historical information clearly labelled with time stamps, and make the LLM agents able to recall relevant information precisely. Let us assume that this can be achieved flawlessly using either long-context LLMs (such as Gemini[48]) or RAG (retrieval augmented generation[49]), although such an approach is still an open challenge in reality. In the meantime, we need to explicitly prompt the LLM agents to use their historical information for decision-making. For instance, the prompt to elicit an LLM agent's decision might include words such as "This is the n-th year since the new technology was released. You should decide whether to adopt the innovation based on your historical conversations and/or advertisements you received, if they exist."

Taking a step back, it seems more convenient if we borrow some merits from rule-based approaches and stop pursuing "pure emergence". For example, we can set a rule for the agents: if you are exposed to the advertisements six times, you must adopt the new product. Implementing this rule using code can be even easier. However, this level of intervention comes at the cost of abandoning everything that has been built around natural language, regardless of indispensable linguistic information, such as agent personas, historical conversations, and self-reflections, although agent operations related to this information also involve rules. There is certainly some subtler way to blend natural language and rules. But it is challenging to empirically/theoretically justify when, where, why, and how to use rule-text mixtures, which is very likely to result in Frankenstein's creature rather than Battle Angel Alita – both are hybrid beings in science fiction, but are perceived very differently in society. LLM-driven AI assistants (such as Manus[50]) may rely on very specific rules regarding context engineering, which defines how an LLM should call tools, retrieve knowledge, organise memories, and make plans to effectively automate people's daily work. Nevertheless, it should be noted that such agents are task-oriented, and their value is determined by their task performance rather than how well they can mimic heterogeneous human individuals in social contexts.

To proceed, we can either assume that we have successfully built purely narrative-driven agents or exceptionally perfect rule-text hybrids. In either case, we are forced to intervene in agent actions more deeply at the cost of further narrowing the scope for agent decisions to emerge spontaneously.

Dilemma IV: "Do you want a different me or not?"—robust role alignment vs. long-term evolution

The LLM agents' memory mechanism serves as a fundamental basis for role evolution, which is crucial for their attitude change towards the new technology. Prior to handling their role evolution, we must ensure the LLM agents indeed play the roles as expected.

For rigorous social simulation, LLM agents should be carefully prompted or fine-tuned to manifest a certain type of economic preference, political leaning, and personality. Researchers have invested considerable effort in aligning LLMs with human values or specific roles[29,51–54]. A recent paper even proposes "LLM psychometrics" to evaluate the "psychology" of LLMs, similar to how humans are studied by psychologists[55]. The challenge in role alignment is to not only ensure LLM outputs exhibit the attributes consistent with what we expect, but also robust enough to persist when faced with various requests[29] or even so-called "jailbreak" techniques[56] – methods that aim to intentionally break their aligned behaviours. For example, the LLM agent – Stony, designed to be a risk-averse traditionalist – might suddenly adopt an unproven new technology after hearing a single positive anecdote from Novy, such as "*I tried it last week and it worked great! And you also saw how amazing it was! Remember?*" Even though Stony does not have relevant memory, the suggestion creates a false sense of shared experience. It is suspicious if Stony suddenly adopts this technology, as this reflects the fragility in role alignment and failure to be alert to fabricated cues.

Again, let us imagine that we have a perfect technique to make the LLM agent align precisely and robustly with its expected behaviour. Nevertheless, here comes a new dilemma: we need not only role alignment but also credible role evolution. Within a long-term social process, such as innovation diffusion over decades, we definitely expect the agents to adapt in response to the change in social norms, peer pressure, individual reflection, and technological progress. For example, Stony should not maintain a rejection attitude towards new technologies persistently but should evolve in ways that are conditional and contextually coherent, as real human individuals might do. Such change can be gradual or even radical, but should not be erratic or arbitrary. At least, every change in attitude should be traceable through the text, such as the LLM agent's historical conversations, reflections on previous attitudes, and the reasoning for the current decisions. That means, if role alignment is too "successful", LLM agents' behaviour may remain stagnant, regardless of the signals of social change and accumulated evidence. On the contrary, if they evolve too fast or incoherently, we may risk the collapse of their identities, obtaining erratic agent behaviour divorced from their expected roles.

Therefore, what we truly need is an LLM agent with a robust baseline of role cognition, which can evolve but without running off the rails. The question is whether such a nuanced evolution can really be modelled through a sequence of constrained yearly conversations, which resonates with the first dilemma: within a brief communication, the LLM agents must be able to respect their prescribed roles while properly adjusting their stances, which represents a whole year of cognitive change. As we have assumed, "cognitive compression" is acceptable, and the thought experiment needs to proceed; let us push such compression further. We imagine that we can build a powerful auxiliary cognitive system to enable LLM agents to summarise beliefs, suggest peer pressure, and quote historical conversations plausibly – even though the legitimacy of mimicking such role evolution by manipulating words is unknown, and despite that, our role is seemingly evolving closer towards a script writer.

Dilemma V: "A sea of words, a desert of meaning"—when agent behaviour does not add up

After all the trade-offs and struggles we have gone through, our LLM-based diffusion model is ready. We hit the "Run" button to start the simulation. LLM agents begin to speak. Thousands of them across many iterations generate verbose conversational text logs. These dialogues quickly accumulate and occupy a large portion of the storage, containing rich linguistic details, such as the reasoning of adoption or rejection, the reflection on peer choices, the consideration of social norms and personal values. Within the text logs, we count "1" and "0" to obtain the aggregated results of yearly adoption, which might or might not add up to a typical S-shaped curve indicative of a diffusion process. No matter what the shape is, we are faced with the ultimate dilemma: how do we interpret the curve?

Unlike the rule-based model, which has manipulable parameters with social meanings, the LLM-based version does not have clear experimental

Table 1 | Modelling purposes and where LLM agents may or may not fit

Modelling Purpose	Definition/Goal	LLM Fit	Assessment
Prediction	To forecast future states or behaviours of a system.	Mixed: LLMs are not designed to extrapolate system-level patterns with calibrated precision. In contrast, LLMs can predict individuals' behaviour under well-defined contexts.	LLM agents generate plausible conversations but lack causal grounding or stability for system-level forecasting. Given their large parameter spaces, LLMs are flexible enough to fit individuals' behaviour, especially in finite action spaces[62].
Explanation	To identify and test mechanisms that produce observed outcomes.	Weak: LLM outputs are hard to decompose into causal mechanisms and function through analysis of language, not mechanisms.	Interpretability is low; text output obscures underlying processes.
Description	To replicate or represent observed behaviour or phenomena.	Moderate: LLMs can simulate realistic individual behaviours and discourse.	Good for micro-level mimicry, but may lack macro fidelity or reproducibility.
Theoretical Exposition	To explore the consequences of hypothetical assumptions or mechanisms.	Mixed: LLMs can explore "what-if" scenarios but may introduce hidden biases.	Useful for narrative experimentation, but not theory-grounded unless tightly constrained.
Illustration	To demonstrate how a model or idea might work using a simplified or intuitive setup.	Mixed: LLMs can effectively illustrate local agent interactions, but struggle to demonstrate clear system-level emergence.	Strong for showing interpersonal dynamics and discourse; weak when the goal is to trace how individual behaviours lead to macro outcomes—unless heavily simplified.
Analogy	To compare one system to another by highlighting structural or behavioural similarity.	Moderate: LLMs can generate analogies but often lack rigour in structural mapping.	Good for suggesting metaphors; less reliable for deep comparative modelling.
Social Learning	To improve shared understanding among a group of people.	Mixed: This can be a fundamental strength of LLM agents, but it needs caution when dealing with trust- and consensus-building.	LLM agents are socially interactive, which is a strength that may support training, role play, participatory modelling, and studying discursive dynamics. However, further exploration is required to sort out whether and how LLM agents can be helpful in situations where trust and consensus are of central importance[63].

levers. Its outputs rely on prompt design, conversational context, and the invisible dynamics within the LLMs. If the curve is too fast or too slow, too flat or too ragged, we cannot simply adjust some parameters and rerun the simulation. Instead, we might have to manually rewrite prompts, modify agent personas, memory mechanisms, or the cognitive system that drives agents to evolve. Even worse, these modifications probably require (directionless) trials and errors, making systematic experiments almost impossible: causal variables cannot be easily identified or isolated; simulation runs can hardly be generalised from one to another; limited quantifiable parameters can be leveraged to calibrate against target datasets or to conduct sensitivity analysis; both agent behaviours and system dynamics are emergent, while their relationship remains opaque.

Consequently, we derive a sea of words, which are linguistically detailed, behaviourally rich, and locally coherent. A crucial question is: Do we care about what the agents say, or about what the system does? If the answer is the former, we have a wealth of text to read, which, however, is unlikely to offer genuine insights into social dynamics. If the answer is the latter, the sea of words is no longer a strength but a liability. Numerous texts become distractions rather than useful clues pointing to causal chains that bridge the gap between the microscopic agent behaviours and macroscopic social patterns. Indeed, identifying causal links is also a key challenge in conventional simulation—LLM agents make this challenge even harder to tackle. That is, the simulation ends up with a desert of meaning, which comes along with increased computational and environmental costs[57–59], convoluted system design, complicated model calibration and validation, and obscured underlying mechanisms of social phenomena. As a result, the plausible behaviours of numerous autonomous agents are closer to a stage play than to a simulation. They generate stories rather than reveal explanations.

A preliminary attempt to position LLMs in simulation

The five dilemmas we discussed above, from temporal mismatch to opaque emergence, are not unique to LLM-based innovation diffusion simulation. They reflect a more general, systemic paradox that appears when LLMs are used as the cognitive engine of social agents in simulation: the pursuit of realism through fine-grained natural language collides with the modelling demand for abstractness. Properly selected realism can empower modelling's utility as an epistemic tool that intentionally simplifies reality, isolates variables, and exposes mechanisms. Logically, realism should not be the top concern of modelling[60]; otherwise, modelling becomes meaningless as the real world itself stands for perfect realism[24], which, however, does not automatically offer answers to how it works unless we ask wisely[61]. This indicates that the LLM agents' high resemblance to humans in terms of expressiveness may obscure the social processes we intend to explore. As we attempted to balance between realism and feasibility throughout the thought experiment, we unavoidably imposed artificial constraints in each modelling decision. These decisions progressively decrease the behavioural naturalness of LLM agents, but do not achieve the clarity and traceability of rule-based abstraction. Eventually, what we derived from the modelling process is not a methodological harmony but an epistemic uncanny valley: the resultant LLM agents appear more realistic than rule-based agents but are recognisably unhuman.

It should be noted that different modelling purposes often have distinct requirements, in terms of abstraction, calibration, validation, etc. The five dilemmas we identified do not point to any inherent shortcomings of LLMs. Instead, they demonstrate the mismatch between a powerful, flexible tool and the purpose of understanding how microscopic mechanisms can generate system-level patterns – a common modelling purpose in human system simulation. In practice, LLM agents have been found to excel in diverse domains in social simulation, where these dilemmas are resolved, mitigated, or even reversed in meaning. ref. 21 summarises seven types of modelling purposes, which can serve as a tentative guide for when we should use LLM-driven simulation. Table 1 briefly estimates where LLM agents may or may not fit these purposes.

Furthermore, the thought experiment has shown that LLMs are not too weak for simulation, but rather, they are often too strong, too vivid, and too dependent on detailed context, which makes them unsuited for the required simplification and abstraction of many existing modelling purposes. Their capability of narrative generation exceeds the level of detail that typical models can accommodate. Hence, LLM agents should not replace

Table 2 | Several "comfort zones" for LLM agents

Comfort Zone	Research Focus	Application Cases
Deliberative Discourse & Argument Framing	Discursive behaviour, rhetorical framing, micro-level argument dynamics	Mock trials[64], stakeholder negotiations[65], persuasion[66,67]
Education, Training, Situated Role Play	Real-time conversation, role-play fidelity, situated dialogue	Classroom simulations[68], empathic conversation[69], clinical role-play[70,71]
Human-in-the-Loop Systems & Speculative Prototyping	Co-creative dialogue, human-guided iteration	Design fiction[72], speculative prototyping[73], decision support systems[74]

conventional simulation approaches but instead extend the landscape, especially in domains that require high linguistic plausibility. LLM agents are ideally suited where system-level emergence is not the focus, where linguistic nuances and meaning are central, where interactions unfold in natural time, and where stable role identity is more important than long-term behavioural evolution. These conditions delineate the "comfort zones" for LLM agents. Table 2 outlines several such comfort zones, not as formal modelling purposes but research or application contexts, where LLM agents can be superior to established rule-based alternatives.

Although we seek to provide clear critical reflections on the use of LLMs in simulation, it would be shortsighted to make definitive conclusions with respect to the role of LLMs. As LLM technologies are still rapidly evolving, new algorithms, training methods, and agent engineering techniques are likely to continuously reshape our understanding of LLM-based simulation. These modelling purposes and comfort zones are neither mutually exclusive, exhaustive nor static. With the development of LLM agents, the fitness of LLMs for these purposes can also vary. The categorisation of modelling purposes needs revision or extension in future research. With this in mind, we suggest that LLM-based modelling is a new avenue for research in social simulation rather than a technological advance within established modelling approaches, but one that is still in need of critical exploration.

Data availability

No data associated with this article.

Received: 28 September 2025; Accepted: 30 January 2026;
Published online: 02 March 2026

References

1. Achiam, J. et al. Gpt-4 technical report. Preprint at https://doi.org/10.48550/arXiv.2303.08774 (2023).
2. Claude. Meet Claude, your thinking partner. https://www.anthropic.com/claude (2025).
3. Cui, Z., Li, N. & Zhou, H. A large-scale replication of scenario-based experiments in psychology and management using large language models. *Nat. Comput. Sci.* https://doi.org/10.1038/s43588-025-00840-7 (2025).
4. Wang, L. et al. A survey on large language model based autonomous agents. *Front. Comput. Sci.* **18**, 186345. https://doi.org/10.1007/s11704-024-40231-1 (2024).
5. Zhang, Z. et al. A survey on the memory mechanism of large language model-based agents. *ACM. Trans. Inf. Syst.* **43**, 1–47, https://doi.org/10.1145/3748302 (2025).
6. Li, X., Wang, S., Zeng, S., Wu, Y. & Yang, Y. A survey on LLM-based multi-agent systems: workflow, infrastructure, and challenges. *Vicinagearth* **1**, 9 (2024).
7. Aoki, N., Mori, N. & OKada, M. Analysis of LLM-based narrative generation using the agent-based simulation. In *2023 15th International Congress on Advanced Applied Informatics Winter (IIAI-AAI-Winter)* 284–289 (IEEE, 2023).
8. Wang, Y., Zhou, Q. & Ledo, D. StoryVerse: Towards co-authoring dynamic plot with LLM-based character simulation via narrative planning. In *Proc. 19th International Conference on the Foundations of Digital Games* 1-4 (ACM, 2024).
9. Yan, Z. & Xiang, Y. Social life simulation for non-cognitive skills learning. *In Proc. ACM Hum.Comput. Interact.* **9**, 1–44 (2025).
10. Horiguchi, I., Yoshida, T. & Ikegami, T. Evolution of social norms in LLM agents using natural language. Preprint at https://doi.org/10.48550/arXiv.2409.00993 (2024).
11. Carrasco-Farre, C. Large language models are as persuasive as humans, but how? About the cognitive effort and moral-emotional language of LLM arguments. Preprint at https://doi.org/10.48550/arXiv.2404.09329 (2024).
12. Li, J. et al. Banishing LLM hallucinations requires rethinking generalization. Preprint at https://doi.org/10.48550/arXiv.2406.17642 (2024).
13. Perković, G., Drobnjak, A. & Botički, I. Hallucinations in llms: understanding and addressing challenges. In 2024 *47th MIPRO ICT and Electronics Convention (MIPRO)* 2084–2088 (IEEE, 2024).
14. Tonmoy, S. et al. A comprehensive survey of hallucination mitigation techniques in large language models. Preprint at https://doi.org/10.48550/arXiv.2401.01313 (2024).
15. Gallegos, I. O. et al. Bias and fairness in large language models: a survey. *Comput. Linguist.* **50**, 1097–1179 (2024).
16. Hu, T. et al. Generative language models exhibit social identity biases. *Nat. Comput. Sci.* **5**, 65–75 (2025).
17. Lin, L., Wang, L., Guo, J., & Wong, K.-F. Investigating bias in llm-based bias detection: Disparities between llms and human perception. In Proceedings of the 31st International Conference on Computational Linguistics,10634-10649. https://aclanthology.org/2025.coling-main.709/ (2025).
18. Tao, Y., Viberg, O., Baker, R. S. & Kizilcec, R. F. Cultural bias and cultural alignment of large language models. *PNAS Nexus* **3**, 346, https://doi.org/10.1093/pnasnexus/pgae346 (2024).
19. Deng, Z. et al. Ai agents under threat: a survey of key security challenges and future pathways. *ACM Comput. Surv.* **57**, 1–36 (2025).
20. Friha, O. et al. Llm-based edge intelligence: a comprehensive survey on architectures, applications, security and trustworthiness. *IEEE Open J. Commun. Soc.* **5**, 5799–5856 (2024).
21. Edmonds, B. Different modelling purposes. *Simulating social complexity: A handbook*, 39–58 (2017).
22. Epstein, J. M. *Generative social science: Studies in agent-based computational modeling* (Princeton University Press, 2012).
23. Knuuttila, T. Epistemic artifacts and the modal dimension of modeling. *Eur. J. Philos. Sci.* **11**, 65 (2021).
24. Miller, J. H. & Page, S. E. *Complex adaptive systems: an introduction to computational models of social life* (Princeton University Press, 2009).
25. Morgan, M. S. *The world in the model: How economists work and think* (Cambridge University Press, 2012).
26. North, M. J. & Macal, C. M. *Managing business complexity: discovering strategic solutions with agent-based modeling and simulation* (Oxford University Press, 2007).
27. Cuskley, C., Loreto, V. & Kirby, S. A social approach to rule dynamics using an agent-based model. *Top. Cognit. Sci.* **10**, 745–758 (2018).

28. Epstein, J. M. & Axtell, R. *Growing artificial societies: social science from the bottom up* (Brookings Institution Press, 1996).
29. Xie, Q. et al. Human Simulacra: Benchmarking the Personification of Large Language Models. Preprint at https://doi.org/10.48550/arXiv.2402.18180 (2024).
30. Zhang, N., Lu, Y., Chen, J. & Hwang, B.-G. An agent-based diffusion model for residential photovoltaic deployment in Singapore: Perspective of consumers' behaviour. *J. Clean. Prod.* **367**, 132793. https://doi.org/10.1016/j.jclepro.2022.132793 (2022).
31. Bass, F. M. A new product growth for model consumer durables. *Manag. Sci.* **15**, 215–227 (1969).
32. Rao, K. U. & Kishore, V. V. N. A review of technology diffusion models with special reference to renewable energy technologies. *Renew. Sustain. energy Rev.* **14**, 1070–1078 (2010).
33. Rogers, E. M., Singhal, A. & Quinlan, M. M. in *An Integrated Approach to Communication Theory and Research* (Routledge, 2014).
34. Bass, F. M. Comments on "a new product growth for model consumer durables the bass model. *Manag. Sci.* **50**, 1833–1840 (2004).
35. Rand, W. & Rust, R. T. Agent-based modeling in marketing: Guidelines for rigor. *Int. J. Res. Mark.* **28**, 181–193 (2011).
36. Alderete Peralta, A. Spatio-temporal modelling of diffusion of electric vehicles and solar photovoltaic panels: an integrated agent-based and artificial neural networks model. Cranfield University (2020).
37. Caprioli, C., Bottero, M. & De Angelis, E. Supporting policy design for the diffusion of cleaner technologies: a spatial empirical agent-based model. *ISPRS Int. J. GEO-Inf.* **9**, 581 (2020).
38. Chen, Z. An agent-based model for information diffusion over online social networks. *Pap. Appl. Geogr.* **5**, 77–97 (2019).
39. El-Sayed, A. M., Scarborough, P., Seemann, L. & Galea, S. Social network analysis and agent-based modeling in social epidemiology. *Epidemiol. Perspect. Innov.* **9**, 1–9 (2012).
40. Ghoulmie, F., Cont, R. & Nadal, J. P. Heterogeneity and feedback in an agent-based market model. *J. Phys.: Condens. matter* **17**, S1259 (2005).
41. Olukan, D. Heterogeneity in Agent-based models, University of Leeds, (2023).
42. Shi, Y., Wei, Z., Shahbaz, M. & Zeng, Y. Exploring the dynamics of low-carbon technology diffusion among enterprises: An evolutionary game model on a two-level heterogeneous social network. *Energy Econ.* **101**, 105399. https://doi.org/10.1016/j.eneco.2021.105399 (2021).
43. Shi, Y. et al. Leveraging inter-firm influence in the diffusion of energy efficiency technologies: An agent-based model. *Appl. Energy* **263**, 114641. https://doi.org/10.1016/j.apenergy.2020.114641 (2020).
44. Zeng, Y., Dong, P., Shi, Y., Wang, L. & Li, Y. Analyzing the co-evolution of green technology diffusion and consumers' pro-environmental attitudes: an agent-based model. *J. Clean. Prod.* **256**, 120384. https://doi.org/10.1016/j.jclepro.2020.120384 (2020).
45. Zeng, Y., Shi, Y., Shahbaz, M. & Liu, Q. Scenario-based policy representative exploration: a novel approach to analyzing policy portfolios and its application to low-carbon energy diffusion. *Energy* **296**, 131202. https://doi.org/10.1016/j.energy.2024.131202 (2024).
46. Kiesling, E., Günther, M., Stummer, C. & Wakolbinger, L. M. Agent-based simulation of innovation diffusion: a review. *Cent. Eur. J. Oper. Res.* **20**, 183–230 (2012).
47. Jones, C.R. & Bergen, B.K. Review: Large Language Models Pass the Turing Test. SuperIntelligence-Robotics -Safety & Alignment 2(2) https://doi.org/10.70777/si.v2i2.14697 (2025).
48. Gemini. Gemini https://deepmind.google/models/gemini/ (2025).
49. Arslan, M., Ghanem, H., Munawar, S. & Cruz, C. A Survey on RAG with LLMs. *Procedia Comput. Sci.* **246**, 3781–3790 (2024).
50. Manus. https://manus.im/ (2025).
51. Feng, Q. et al. EmoCharacter: evaluating the emotional fidelity of role-playing agents in dialogues. In *Proc. 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* 6218–6240 (Association for Computational Linguistics, 2025).
52. Liu, Y. et al. Trustworthy llms: a survey and guideline for evaluating large language models' alignment. Preprint at https://doi.org/10.48550/arXiv.2308.05374 (2023).
53. Yu, T. et al. Aligning multimodal llm with human preference: a survey. Preprint at https://doi.org/10.48550/arXiv.2503.14504 (2025).
54. Zhou, D., Zhang, J., Feng, T. & Sun, Y. A survey on alignment for large language model agents. In *UIUC Spring 2025 CS598 LLM Agent Workshop* (2025).
55. Ye, H., Jin, J., Xie, Y., Zhang, X. & Song, G. Large Language Model Psychometrics: A Systematic Review of Evaluation, Validation, and Enhancement. Preprint at https://doi.org/10.48550/arXiv.2505.08245 (2025).
56. Wei, A., Haghtalab, N. & Steinhardt, J. Jailbroken: how does llm safety training fail? *Adv. Neural Inf. Process. Syst.* **36**, 80079–80110 (2023).
57. Büchel, J. et al. Efficient scaling of large language models with mixture of experts and 3D analog in-memory computing. *Nat. Comput. Sci.* **5**, 13–26 (2025).
58. Jiang, P., Sonne, C., Li, W., You, F. & You, S. Preventing the immense increase in the life-cycle energy and carbon footprints of llm-powered intelligent chatbots. *Engineering* **40**, 202–210 (2024).
59. Rillig, M. C., Ågerstrand, M., Bi, M., Gould, K. A. & Sauerland, U. Risks and benefits of large language models for the environment. *Environ. Sci. Technol.* **57**, 3464–3466 (2023).
60. Kokko, H. *Modelling for field biologists and other interesting people* (Cambridge University Press, 2007).
61. Kant, I. *Critique of pure reason* (Minerva Heritage Press, 2024).
62. Grote, T., Freiesleben, T. & Berens, P. Foundation models in healthcare require rethinking reliability. *Nat. Mach. Intell.* **6**, 1421–1423 (2024).
63. Kolagani, N. et al. Participatory modeling in the AI era. *Environ. Model. Softw.* **196**, 106762. https://doi.org/10.1016/j.envsoft.2025.106762 (2025).
64. He, Z. et al. AgentsCourt: building judicial decision-making agents with court debate simulation and legal knowledge augmentation. Preprint at https://doi.org/10.48550/arXiv.2403.02959 (2024).
65. Abdelnabi, S., Gomaa, A., Sivaprasad, S., Schönherr, L. & Fritz, M. Cooperation, competition, and maliciousness: LLM-stakeholders interactive negotiation. Preprint at https://doi.org/10.48550/arXiv.2309.17234 (2024).
66. Hackenburg, K., Ibrahim, L., Tappin, B.M. & Tsakiris, M. Comparing the persuasiveness of role-playing large language models and human experts on polarized U.S. political issues. *AI & Soc* **41**, 351–361, https://doi.org/10.1007/s00146-025-02464-x (2026).
67. Rogiers, A., Noels, S., Buyl, M. & De Bie, T. Persuasion with large language models: a Survey. Preprint at https://doi.org/10.48550/arXiv.2411.06837 (2024).
68. Zhang, Z. et al. Simulating classroom education with llm-empowered agents. *In Proc. 2025 Computational Linguistics: Human Language Technologies* **1**, 10364–10379, https://doi.org/10.18653/v1/2025.naacl-long.520 (2025).
69. Wang, Y., Zhang, H. & Shang, J. EmpLLM: enhancing Empathy in LLMs through psychologist simulation. in *International Conference on Intelligent Multilingual Information Processing* 205–218 (Springer).
70. Li, Y. et al. Leveraging large language model as simulated patients for clinical education. Preprint at https://doi.org/10.48550/arXiv.2404.13066 (2024).
71. Louie, R. et al. Roleplay-doh: Enabling Domain-Experts to Create LLM-simulated Patients via Eliciting and Adhering to Principles. In *Proc. 2024 Conference on Empirical Methods in Natural Language Processing,* 10570-10603. https://doi.org/10.18653/v1/2024.emnlp-main.591 (2024).

72. Wu, H., Wu, W., Xu, T., Zhang, J. & Zhao, H. Towards enhanced immersion and agency for LLM-based interactive drama. Preprint at https://doi.org/10.48550/arXiv.2502.17878 (2025).
73. Tost, J. et al. Futuring machines: an interactive framework for participative futuring through human-ai collaborative speculative fiction writing. In *Proc. 6th ACM Conference on Conversational User Interfaces* 1–7 (ACM, 2024).
74. Li, Z., Zhu, H., Lu, Z., Xiao, Z. & Yin, M. From text to trust: empowering AI-assisted decision making with adaptive LLM-powered Analysis. In *Proc. 2025 CHI Conference on Human Factors in Computing Systems* 1–18 (ACM, 2025).

Acknowledgements

This work was supported by the Helmholtz Excellence Recruiting Initiative, Climate Mitigation and Bioeconomy Pathways for Sustainable Forestry (CLIMB-FOREST; grant no. 101059888) and Co-designing Holistic Forest-based Policy Pathways for Climate Change Mitigation (grant no. 101056755) projects.

Author contributions

Y.Z. designed the work, developed the conceptual foundation and thought experiment, wrote the original draft and prepared the figure and tables. C.B. and M.R. reviewed and edited the manuscript, managed the project administration. All authors have read and approved the final manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Yongchao Zeng.