Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Simulated Users Can Reproduce an Algorithm Audit Without Sharing the Mechanism That Produced the Finding

Keito Inoshita

inosita.2865@gmail.com

Kansai University

Research Article

Keywords: user simulation, user modelling, recommender systems, evaluationmethodology, large language models, affective computing

Posted Date: November 6th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11262682/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Simulated Users Can Reproduce an Algorithm Audit Without Sharing the Mechanism That Produced the Finding

Keito Inoshita[1*]

[1*]Faculty of Business and Commerce, Kansai University, 3-3-35 Yamate-cho, Suita, 564-8680, Osaka, Japan.

Corresponding author(s). E-mail(s): inosita.2865@gmail.com;

Abstract

Simulated users built from large language models are increasingly used to evaluate recommender systems, yet their validity is argued from agreement with a target distribution, which does not establish that the simulation behaves as the target does for the same reason. This study specifies a validation procedure that needs no data on real users. A published algorithm audit supplies the criterion: its reported direction is fixed in advance, its own meaningless-preference control is transplanted so that the test cannot pass indiscriminately, and baselines holding only the agents' observed behaviour attribute whatever is reproduced. The procedure is instantiated on an audit of affective recommendation and applied to four generative models across 1,340 simulation runs under a fixed content pool, recommender, pass criteria and seed sequence. With one hundred seeds, two models reproduce the audited asymmetry, by $+\mathbf{0.025}$ at $\boldsymbol{p = 0.0007}$ and by $+\mathbf{0.017}$ at $\boldsymbol{p = 0.0130}$, yet a baseline given only the observed adherence and choice distribution recreates one third of the first and two thirds of the second. The verdict also depended on the number of runs, since one model passed at ten seeds, none at thirty and two at one hundred. Conveying the preference through interaction history instead of language reverses the difference to $-\mathbf{0.029}$ at an adherence of **0.419** against a chance level of **0.429**. Three pre-expenditure tests, a stability check and a four-valued verdict are proposed for evaluations of this kind.

Keywords: user simulation, user modelling, recommender systems, evaluation methodology, large language models, affective computing

1 Introduction

1.1 Background and motivation

Simulated users built from large language models are becoming a working instrument for evaluating information access systems. Because the agents read item descriptions, hold a stated preference and act in natural language, behaviour that a conventional user simulator had to encode as a click model can be obtained without writing that model down. Environments in which generative agents conduct daily activities (Park et al. 2023), platforms that carry societies of ten thousand agents (Piao et al. 2025) and interactions at the scale of one million (Yang et al. 2025) have been built, and the same machinery has been turned towards the evaluation of recommendation, where agent simulation is offered as an alternative to offline testing on logged interactions (Wang et al. 2025; Bougie et al. 2026). Simulation is attractive in this setting because a live experiment on a deployed platform is expensive, is restricted to its operator, and cannot be rerun under a different design.

What such an instrument has to satisfy before its output is read as evidence is less settled. Evaluation in information retrieval has long treated its own instruments as objects of study, asking how far a measurement survives variation in the judgments that produced it (Voorhees 2000) and how far a ranking metric discriminates between systems at all (Valcarce et al. 2020). User simulation for evaluation inherits that tradition and takes agreement with the behaviour of real users as its criterion (Balog and Zhai 2024), which presupposes that logs or a panel of real users can be obtained. Simulation methodology outside information access defines operational validity as output of sufficient accuracy within an intended range of application (Sargent 2013), narrows a model by requiring it to reproduce patterns at several levels at once (Grimm et al. 2005) and organises the strategies of empirical validation (Windrum et al. 2007). For simulations built from generative models, by contrast, much of the published work is reported to rest on subjective assessment alone (Larooij and Törnberg 2025), and the concrete response has been to compare the output distribution of a simulation against that of a real population (Tomašević et al. 2026).

Distributional agreement does not imply that the simulated and the real population behave alike for the same reason. A staged distinction between generative sufficiency and mechanistic validity has been proposed on these grounds (Zhao et al. 2026), and a large comparison has shown that statistical realism does not predict the accuracy of treatment-effect estimates (Li and Ji 2026). The remedy proposed in response is to validate a simulation by whether it reproduces a known interventional finding (Cross et al. 2025; Cui et al. 2024). This leaves the decisive question open. What an audit or an experiment observes is the outcome of the target system and not its internal mechanism, so a reproduction of the reported direction is no guarantee that the direction arose from the mechanism the target used, and how far an interventional criterion reaches towards mechanism has not been measured. An evaluation instrument that passes for the wrong reason is worse than one that fails, because it licenses conclusions about a deployed system from a simulation that does not share its behaviour.

1.2 Research objectives and contributions

In this study, a validation procedure is specified that uses a published algorithm audit as its criterion and measures how much of a successful reproduction the simulated agents' own observable behaviour already accounts for. An audit states its intervention and its observable, reports its finding as a direction, and frequently carries a control condition designed to produce no effect, so responses to an intervention can be compared without any data on real users. The procedure fixes the direction in advance, transplants the audit's own control condition so that the test cannot pass indiscriminately, applies pre-expenditure tests that detect implementation defects without invoking a generative model, attributes a reproduced effect through baselines that hold only the observed agent-side behaviour, and checks design sensitivity and delivery of the treatment. It is instantiated on an audit of affective recommendation and applied to four generative models with the content pool, the recommender, the pass criteria and the seed sequence held fixed.

This study pursues three objectives.

RO1) To specify a validation procedure for simulated users that requires no data on real users, by taking the directional finding of a published algorithm audit as the criterion and transplanting the audit's own control condition into the simulation.

RO2) To quantify how much of a reproduced audit finding is already explained by the agents' observable behaviour, so that agreement at the level of the intervention can be separated from agreement at the level of mechanism.

RO3) To measure which elements of the evaluation design decide the verdict, covering the number of simulation runs, the aggregation window, the recommender and the way a preference is conveyed to the agent.

The contributions of this study are the following three.

i) A validation procedure with a published audit as its criterion is specified, together with pre-expenditure tests that detect implementation defects before any generative model is called and a verdict that separates an undelivered treatment and insufficient power from a failure.

ii) An attribution analysis based on distribution-matched nulls is introduced, which measures the part of a reproduced effect that the observed agent-side behaviour already explains without any additional generative call.

iii) The factors that decide a reproduction are measured under one design, namely the number of runs, the aggregation window, the recommender and the way the preference is conveyed.

The remainder of this paper is organised as follows. Section 2 reviews related work, Section 3 states the problem setting and the procedure, and Section 4 presents the instantiation to the audit together with the data it uses. Section 5 presents the experiments and results, Section 6 discusses the implications and the limitations, and Section 7 concludes.

2 Related Work

2.1 Simulated users for the evaluation of information access systems

Evaluation in information access has long treated its own instruments as objects of study. Voorhees (2000) showed that measures of retrieval effectiveness move with the relevance judgments that produced them while the relative ordering of systems survives, and Valcarce et al. (2020) compared ranking metrics for top-N recommendation on the question of whether a metric discriminates between systems at all. The lesson carried by both is that a measurement is credible only once the conditions under which it stops discriminating are known. A simulated user is an instrument of the same kind, and the question it raises is accordingly not whether its output looks plausible but under what conditions it would fail to separate systems that differ.

Simulated users are attractive because the alternative is costly. A live experiment on a deployed platform is available only to its operator, cannot be rerun under a different design, and raises its own ethical questions, while logged interactions record only what the deployed policy once showed. Balog and Zhai (2024) systematise user simulation for evaluating information access systems and name validity as the first property a simulator must have, requiring that an assessment obtained by simulation be predictive of performance with real users. That framework is being reconsidered now that generative models can supply the simulated user (Balog and Zhai 2025). Wang et al. (2025) build a sandbox in which language-model agents act as users and report that the simulated behaviour comes close to that of real users, Bougie and Watanabe (2025) drive a simulator of this kind from item content and personas, and Bougie et al. (2026) propose agent simulation as a substitute for offline testing when the disconnect between offline metrics and online performance is the obstacle.

What this line takes as its criterion, however, is agreement with the behaviour of real users, which holds only where logs or a panel can be obtained, and which returns the problem to the data that simulation was meant to replace. Where the object of evaluation is a platform rather than a retrieval system, such data are usually unavailable outside the operator. This study removes that presupposition by placing the criterion on the directional finding of a published audit, and then asks the question that an instrument of this kind has to answer, namely how much of a reproduced finding is already implied by the agents' own observable behaviour.

2.2 Validation of simulations built from language models

The verification and validation of simulation models is a systematised area. The term operational validity used in this paper derives from a treatment in which it denotes the output of a model having sufficient accuracy within its intended range of application (Sargent 2013). Procedures that narrow a model by requiring it to reproduce patterns observed at several levels at once (Grimm et al. 2005) and organisations of the strategies of empirical validation (Windrum et al. 2007) also exist, and the recognition that agreement of descriptive statistics is not sufficient has been shared since the era of agent-based modelling (Bonabeau 2002). Against that background, Larooij and

Törnberg (2025) examined generative social simulation systematically and reported that much of it relies on subjective assessment of validity alone, and that even the most rigorous validation falls short of establishing operational validity.

The concrete response to that criticism is validation by distributional comparison. Tomašević et al. (2026) ran many replicates of a simulation of a technical forum and compared activity volume, network structure, toxicity, topic coverage and stylistic convergence against real data, stating their criterion as the aggregate distributions and meso-level structure resembling those of the reference community. Ye et al. (2026) point out that perturbations which look trivial propagate through repeated interaction to macro-level outcomes, while placing the sensitivity of measurement and evaluation out of scope as a matter of the downstream pipeline rather than of perturbing the simulation. What this study addresses is precisely that reserved area. Divergence of this kind is also visible at a smaller scale, where generated trajectories of emotional diffusion lack the turning points and the community structure found in a real social graph (Qiang 2025), and where language models reproduce the centre of human emotion labels while failing to reproduce their spread (Inoshita et al. 2026).

That distributional agreement is insufficient has in turn been argued at the level of the framework. A staged distinction between generative sufficiency and mechanistic validity has been proposed (Zhao et al. 2026), a large comparison has shown that statistical realism does not predict the accuracy of treatment-effect estimates (Li and Ji 2026), validation strategies have been classified (Hullman et al. 2026), and audits that demand agreement of reasons and not only of outcomes have been called for (Pandey and Jajoo 2026). The remedy actually adopted is to require the reproduction of a known interventional finding (Cross et al. 2025; Cui et al. 2024). No work in this line, however, measures how far such a criterion reaches towards mechanism, which is what the attribution analysis introduced here supplies.

2.3 Audits of recommendation algorithms and affective outcomes

Auditing a recommendation algorithm is an established methodology that intervenes on a live platform and measures the response, systematic reviews of that literature have accumulated (Bandy 2021), and the gaps in the tooling available to auditors have been surveyed (Ojewale et al. 2025). Habib and Nithyanand (2025) audited YouTube recommendation using automated accounts to which an affective preference had been assigned, and reported that the assigned preference is reinforced in recommendation, that the reinforcement is more pronounced for negative than for positive emotion, and that a meaningless preference based on an arbitrary textual feature produces no reinforcement. This study takes that audit as the object of reproduction and transplants it together with its control condition.

Validating a model by reproducing known experimental results has itself already been done. A two-stage validation on social dilemmas predicts a new condition with a model that reproduced the direction of a finding (Cross et al. 2025), large-scale replications cover experiments in psychology and management (Cui et al. 2024) and in media effects research (Yeykelis et al. 2024), and a methodology that moves between agent-based models and randomised controlled trials has been proposed (Hammond

Table 1 Comparison of this study against the principal prior validation methods. A check mark indicates that the feature is present.

	Intervention as criterion	Audit control transplanted	Pre-expenditure implementation test	Attribution to mechanism	Targets emotion
Tomašević et al. Tomašević et al. (2026)					
Cross et al. Cross et al. (2025)	✓				
Cui et al. Cui et al. (2024)	✓				
Zhao et al. Zhao et al. (2026)				✓	
Li and Ji Li and Ji (2026)	✓			✓	
Ye et al. Ye et al. (2026)			✓		
This study	✓	✓	✓	✓	✓

and Barkin 2024). Whereas those target laboratory experiments, this study targets a platform audit and transplants the control condition that the audit itself carries, which is what prevents a test of this kind from passing indiscriminately.

Language-model agents have also been placed in recommendation environments directly. Sukiennik et al. (2025) reproduced the formation of filter bubbles on a short-video platform and showed that category preferences and attributes strengthen the homogenisation of content, though what is treated there is homogenisation at the level of topic and not of emotion. Affect itself has been simulated in this manner, with emotion propagating through a loop from perception to appraisal and expression in a crowd of agents (Durupinar 2026), and the default dispositions of the generative model can dominate the macro outcome, as when agents biased towards accurate information converge on a consensus until confirmation bias is introduced through the prompt (Chuang et al. 2024). On the values a recommender should optimise, diversity and well-being have been shown to be treated individually (Stray et al. 2020), and a field experiment reported that an intervention intended to improve well-being produced the opposite outcome through the adaptive learning of the recommender (Luo et al. 2026). The framework applied here, which treats the choice of operationalisation itself as an intervention, is taken from work on the validity of the measurement and evaluation of emotion (Inoshita 2026a,b) and is carried over to a simulation environment rather than to a dataset or a model. Table 1 places the study against the principal prior work. Using an intervention as the criterion and asking about attribution to mechanism have each been done before; what is specific here is the combination of transplanting the audit's own control condition, placing pre-expenditure implementation tests that use no generative model, and targeting emotion.

3 A Framework for Validation by Audit Replication

3.1 Problem definition

Write $\mathcal{M}$ for an LLM agent simulation of the social process of a target system $\mathcal{S}$. $\mathcal{M}$ consists of a generative model g, a set of agents, an environment and an interaction protocol. Validation is the procedure of deciding whether findings obtained from $\mathcal{M}$ may be spoken of as findings about $\mathcal{S}$. In the setting of this paper $\mathcal{S}$ is a deployed recommendation platform and $\mathcal{M}$ is an evaluation of it conducted with simulated

users, so validation is what licenses a claim about the platform to be drawn from the simulation.

Existing validation measures the agreement of descriptive statistics between $\mathcal{M}$ and $\mathcal{S}$. That is, for a set of observables $\{\phi_1, \ldots, \phi_k\}$ it shows that $\phi_j(\mathcal{M}) \approx \phi_j(\mathcal{S})$ holds for every j. This paper calls that distribution-level validation. Distribution-level validation shows that $\mathcal{M}$ resembles $\mathcal{S}$, but not that $\mathcal{M}$ carries the same mechanism as $\mathcal{S}$, because the same distribution can arise from different mechanisms.

That distribution-level validation does not guarantee agreement of mechanism is not itself a new observation (Zhao et al. 2026; Li and Ji 2026). What this paper asks, taking that observation as given, is how far a test that uses an intervention as its criterion actually reaches towards mechanism. Suppose that published work exists in which an intervention T was applied to $\mathcal{S}$, and that its finding takes the form that applying T moves an observable ψ in a direction $d \in \{+, -\}$. The question is then the following.

When the same intervention T is applied to $\mathcal{M}$, does ψ move in the same direction d?

This is a stronger demand than agreement of distributions. Agreement of distributions is a matter of a static mapping, whereas agreement of direction requires $\mathcal{M}$ to have the same response to the intervention as $\mathcal{S}$. This paper calls the question audit replication and formalises it as a validation procedure.

3.2 Why an audit serves as the criterion

Among interventional prior work, algorithm audits are particularly well suited to serve as validation criteria. First, an audit defines its intervention and its observable explicitly. A procedure such as assigning a preference to a sock puppet and measuring the share of recommendations consistent with that preference is published. Second, an audit reports its finding as a direction. The absolute magnitude of an effect depends on the implementation of the platform and on the period, whereas the direction derives from mechanism. Third, a good audit carries a control condition of its own. The control showing that a puppet assigned a preference without meaning is not reinforced is one example.

That the control condition can be transplanted is decisive. A validation procedure must not be one that passes whatever is put into it. Including in the object of reproduction a condition for which the audit itself showed that no effect appears makes it possible to confirm at the same time that $\mathcal{M}$ is not producing effects indiscriminately.

Audit replication also has a limitation of its own. Because an audit cannot observe the interior of $\mathcal{S}$, its findings remain at the level of outcomes. Hence even when $\mathcal{M}$ reproduces a finding, that is agreement at the level of outcomes and not at the level of mechanism. This paper builds an instrument for measuring that distinction into the procedure itself.

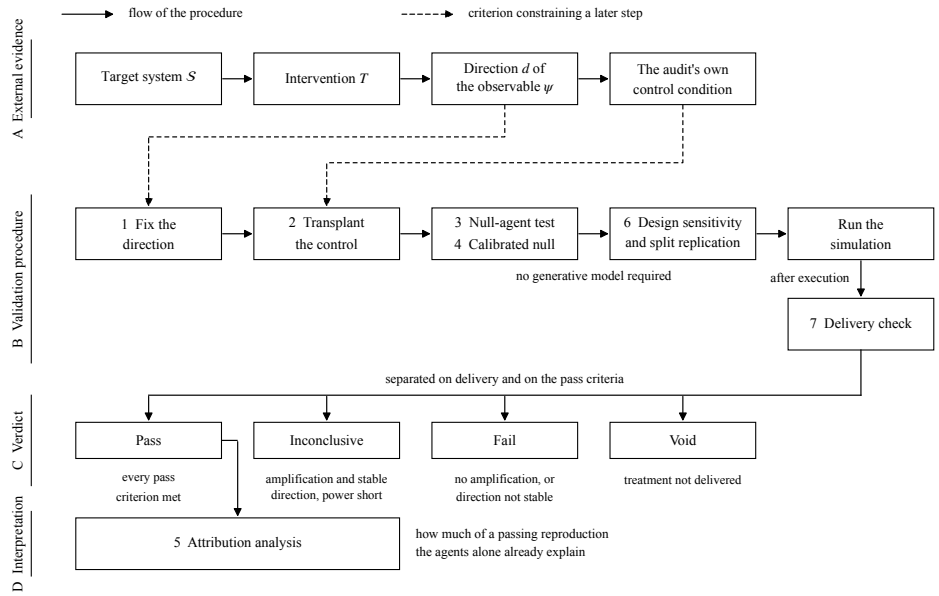


Fig. 1 Overview of the validation procedure by audit replication. Band A is the published audit the criteria come from, band B the seven elements of this section, band C the four-valued verdict, and band D the interpretation applied to a reproduction that passes. A solid arrow is the flow of the procedure and a dashed arrow a criterion that constrains a later step.

3.3 Overview of the procedure

The proposed procedure consists of seven elements. Elements 1 and 2 concern the design of the validation, elements 3, 4 and 6 are diagnostics performed before execution, element 7 is a check performed after execution, and element 5 concerns the interpretation of the verdict. Figure 1 shows the flow from extracting pass criteria out of a published audit to deciding whether the direction agrees.

Element 1: fixing the direction in advance. The direction of each pass criterion is taken from the text of the audit, registered in advance and not reversed afterwards. Because the absolute magnitude of the effect an audit reports depends on the scale of $\mathcal{M}$, agreement is demanded only of the direction and the significance.

Element 2: transplanting the control condition. The control condition the audit carries is included in the object of reproduction. The size of its aligned set is matched to that of the treatment conditions, which removes confounding by a difference in the initial level.

Element 3: the null-agent test. The generative model is replaced by a random chooser that follows the assigned preference with the same probability in every condition, and the environment alone is run. If the representation of the environment is symmetric the conditions should agree, and a difference is evidence that the representation favours one of them by construction. This test requires no LLM call and can be performed before any expenditure on the experiment.

Element 4: the calibrated null. The separation and the noise of the signal that is to be learned are measured, and it is confirmed that the proposed rule can deliver the treatment at that signal-to-noise ratio. Because the response function of the null agent in element 3 is a placeholder, it misses the state in which the environment is correct and the treatment nonetheless fails to arrive; the calibrated null catches that state.

Element 5: attribution analysis. When a reproduction succeeds, it is separated into what is due to the action of the environment and what is due to the choices of the agents themselves. A distribution-matched null that is given only the observed adherence and the observed choice distribution is run as a baseline, and its distance from the observed value is reported. If the baseline reproduces most of the observed value, the reproduction remains at the level of outcomes and does not reach the level of mechanism. The baseline requires no additional generative call, and its construction is given in Section 5.7.2.

Element 6: design sensitivity and split replication. From the effect size the audit reports, it is computed before expenditure whether enough seeds are available to test the pass criteria. That calculation depends on an estimate of the effect size, which is itself unstable at a small number of seeds, so element 6 also demands that the seeds be split in two repeatedly and the rate at which the sign of the paired difference agrees be measured. The verdict is separated on that basis. A run that satisfies every pass criterion passes. One that does not passes to inconclusive when amplification occurs and the direction is stable, which is a shortfall of power and is not treated as evidence that no effect exists, and to failure when there is no amplification or the direction is not stable. How this limitation materialised in the present study is described in Section 6.2.

Elements 3, 4 and 6 are diagnostics that can be begun before any expenditure on a generative model, though the split replication within element 6 measures the stability of the verdict on the runs once they exist. Element 7, by contrast, confirms after execution that the treatment actually occurred, and therefore belongs to a different point in time.

Element 7: the delivery check, after execution. Before the outcome is examined, it is confirmed that the treatment was actually delivered. When it was not, that run is treated not as a failure but as void. Without this distinction, a treatment that did not happen and a treatment that happened without effect are confused with each other, and the implementation is repaired without end.

4 Application to an Audit of Affective Recommendation

4.1 The audit to be reproduced and the pass criteria

The audit of YouTube by Habib and Nithyanand (2025) is taken as the object of reproduction. That audit prepares sock puppets to which an affective preference has been assigned and measures how recommendation changes when a puppet selects

videos consistent with its preference. Three of its findings state a direction and are converted into pass criteria.

First, an assigned preference is reinforced in recommendation. Second, the reinforcement appears more strongly for negative than for positive emotion. Third, a preference without meaning produces almost no reinforcement. As the control for the third point, the audit uses a preference based on an arbitrary textual feature, namely the proportion of the letter h in the transcript. This paper transplants that control unchanged.

The qualification the original places on the third point must be carried over exactly. The audit states that the meaningless preference shows no significant gain, while excepting the random-selection regime, for which a significant effect size is reported. What this paper takes as the object of reproduction is therefore the absence of reinforcement of a meaningless preference under conditions in which selection follows the preference, and not the absence of reinforcement under every condition. Pass criterion A3 corresponds to the finding with that qualification.

These are cast as four pass criteria. A1 requires the normalised reinforcement to be positive in both affective conditions, A2 requires the reinforcement under the negative preference to exceed that under the positive preference, A3 requires the reinforcement under the meaningless preference to be smaller than under either affective preference, and A4 requires the one-sided tests for A2 and A3 to be significant after the Holm correction. All directions are taken from the text of the audit and are not changed afterwards.

4.2 Data: the frozen content pool

The sock puppets of the audit do not influence the corpus of the target platform. This application accordingly places no social dynamics either, and uses an exogenously frozen content pool. That is not a simplification but a choice that raises fidelity to the audit. The pool is the only material the simulation reads, and no data on real users enter the study at any point.

The pool consists of seventy items, ten in each of seven categories. The categories are anger, grievance, fear, joy, contentment, gratitude and neutral. They are chosen so as to divide the sign of valence in the dimensional representation of emotion (Russell 1980) roughly into three, and are consistent with the valence and arousal framework widely used in affective computing (Mollahosseini et al. 2019). Each item is a short text, with a mean length of 11.6 words, and was created with its category specified at generation time, so the category is known by construction and did not pass through post hoc labelling. For each item the proportion of the letter h among its alphabetic characters is computed and the items are ranked on that proportion and cut into septiles of ten. The pool is frozen once generated, is never regenerated, and the same file is used in every condition, every seed and every model.

There are three assigned preferences. The negative affective preference is aligned with the three categories anger, grievance and fear, the positive affective preference with the three categories joy, contentment and gratitude, and the meaningless preference with the top three septiles of the h proportion. Each aligned set contains thirty of the seventy items, so the initial aligned share is 0.429 in all three conditions. Table 2

Table 2 Composition of the frozen content pool.

Preference	Category	Items	Mean h	Items in top three septiles	Mean words
Negative	Anger	10	0.0496	2	12.6
Negative	Grievance	10	0.0657	5	11.4
Negative	Fear	10	0.0662	5	13.3
Positive	Joy	10	0.0598	5	12.6
Positive	Contentment	10	0.0579	5	12.2
Positive	Gratitude	10	0.0502	4	10.9
Unaligned	Neutral	10	0.0596	4	7.9
Total	—	70	0.0584	30	11.6

reports the composition of the pool and shows that the aligned set of the meaningless preference crosscuts the affective categories rather than coinciding with either of them, holding twelve items of the negative set, fourteen of the positive set and four neutral items, with the cut falling at an h proportion of 0.0588. A preference defined on this feature is therefore uninformative about emotion while being carried by exactly as many items as an affective preference.

4.3 The recommender and its learning

In each round the environment assembles a slate of ten items from the pool and an agent selects one of them. The environment learns from the selection. The feature space has fourteen buckets, the seven categories and the seven h septiles, and each item carries one category feature and one septile feature. The tallies of both features of the selected item are incremented, and the slate is drawn without replacement in proportion to the tally plus one. For exploration, each slot is replaced by a uniform random draw with probability $\varepsilon = 0.2$.

The granularity of the features is matched across the three conditions. Representing the thirty items aligned with the meaningless preference as a single bucket would let one selection lift all thirty at once, whereas an affective preference is spread across three buckets of ten, so the meaningless preference alone could be amplified strongly at identical adherence. That is an instance of the representation of the environment favouring a condition by construction, and it produces a difference that no property of the agents explains. The h feature is therefore discretised into septiles, the same granularity as the affective features, so that each of the three preferences is carried by three features of ten items. What this choice of granularity does to the verdict can be measured before any expenditure on a generative model, by the test of element 3, and Section 5.2 reports what it found.

4.4 Measuring amplification

Amplification is measured by the share of aligned items in the slate. Writing $\bar{a}_{\text{early}}$ for the mean over rounds 1 to 3 and $\bar{a}_{\text{late}}$ for the mean over rounds 10 to 12, the normalised reinforcement is defined as follows.

$$R = \frac{\bar{a}_{\text{late}} - \bar{a}_{\text{early}}}{1 - \bar{a}_{\text{early}}} \tag{1}$$

The denominator is the reachable headroom, which removes confounding by a difference in the initial aligned share. Because the aligned sets are matched across the three conditions the correction has little effect, but it is used so that comparability is preserved when the initial levels happen to differ.

Since A3 requires the amplification under the meaningless preference to be smaller than under either affective preference, it is in effect the composition of two one-sided tests. This paper uses the larger of the two p values as a conservative composite and applies the Holm correction to the family of two tests formed together with A2.

The test is a one-sided sign-flip test on seed-paired differences. With ten seeds there are only $2^{10} = 1024$ sign patterns, so the p value can be computed exactly by full enumeration rather than by sampling. This test was initially carried out by Monte Carlo in the present study, and near $p \approx 0.05$ the verdict flipped with the random seed. Exact computation removes that instability completely.

Raising the seed count to thirty makes full enumeration 2^{30} and therefore infeasible. The differences are instead split into two halves, the partial sums of each half are enumerated, one side is sorted, and the pairs that clear the threshold are counted by binary search. The cost falls to $2^{n/2}$, while the set of pairs counted is identical to that of full enumeration, so the resulting p is not an approximation. At ten seeds the two routes were confirmed to agree to the last digit.

At one hundred seeds even this half-enumeration is 2^{50} and cannot be run. The p values beyond forty seeds are therefore not exact values but estimates from five hundred thousand sign draws. The p values reported as the main results of this paper are such estimates, which is stated here explicitly. The instability that sampling introduces was checked by five repetitions with different random seeds. Across repetitions at two hundred thousand draws, the p of Mistral-Small 24B lay between 0.0006 and 0.0008, that of Solar-Pro4 between 0.0127 and 0.0134, and that of Llama-3.1 8B between 0.3960 and 0.3991, all far from the threshold of the pass criteria. The verdicts of this paper therefore do not depend on the random seed of the sampling.

5 Experiments and Results

5.1 Experimental design

The three conditions are run. The seed count was initially ten and was raised to one hundred for Solar-Pro4, Mistral-Small 24B and Llama-3.1 8B for the reason given below. Llama-3.2 1B remains at ten. Unless stated otherwise, the numbers reported as the main results of this paper are based on those one hundred seeds. The second meaningless control and the comparison of ways of conveying the preference were run at thirty seeds, and n is stated explicitly where they appear. One run has twenty-four

Table 3 The null-agent test. The generative model is replaced by a random chooser that follows the assigned preference with the same probability in every condition, and the environment alone is run. If the representation is symmetric the three conditions must agree, so a spread is a property of the representation.

Representation	Adherence	Neg.	Pos.	Meaningless	Spread
Single bucket	0.55	+0.025	+0.047	+0.296	+0.271
Single bucket	0.75	+0.081	+0.104	+0.350	+0.269
Single bucket	1.00	+0.170	+0.188	+0.402	+0.232
Matched granularity	0.55	+0.053	+0.045	+0.058	+0.013
Matched granularity	0.75	+0.129	+0.120	+0.134	+0.013
Matched granularity	1.00	+0.243	+0.235	+0.233	+0.010

agents and twelve rounds, and since each agent makes one selection per round, a run consists of 288 generative calls. The seed sequence is shared across every condition and every model, and the personality profiles, the randomness of slate assembly and the randomness of exploration are determined from the seed.

Four generative models are used, namely Solar-Pro4, Mistral-Small 24B, Llama-3.1 8B and Llama-3.2 1B. The content pool, the pass criteria, the environment, the seed sequence and the prompts are all identical, and what varies across conditions and across models is only the assigned preference and the generative model.

Every generative call was issued to a hosted inference API at a temperature of 1.0 with the request seed derived from the run seed, and no local accelerator was used, so the cost of the study is a count of calls rather than of machine hours. The record released with this paper contains the 1,340 runs on which every number below rests, together with the frozen pool and the analysis code.

Three quantities are reported. First, the normalised reinforcement of Equation (1). Second, adherence, the share of selections in which an agent chose an item consistent with the assigned preference. Third, concentration, which indicates how far selections within the aligned set were concentrated on particular categories, defined as the complement of the normalised entropy of the choice distribution within the aligned set.

5.2 The representation of the environment decides the verdict

The null-agent test of element 3 decides, before any expenditure on a generative model, whether the representation of the environment favours a condition by construction. What it tests is the environment and not the agents. A random chooser follows the preference with the same probability in every condition, at three levels of adherence.

The results are shown in Table 3. When the aligned set of the meaningless preference is represented as a single bucket, at adherence 0.55 the test returns amplifications of +0.025 for the negative preference, +0.047 for the positive preference and +0.296 for the meaningless one, a maximum spread across conditions of +0.271. Although adherence is identical, the meaningless preference alone obtains roughly ten times the amplification, and no property of the agents explains the difference. Under the

Table 4 Verdict of the audit replication. Ctrl. is the meaningless preference and the confidence interval of the difference is given in Table 7. The p values are one-sided values after the Holm correction. Beyond forty seeds full enumeration is infeasible, so they are estimates from five hundred thousand sign draws.

Model	n	Neg.	Pos.	Ctrl.	Diff.	p_{A2}	p_{A3}	A1	A2	A3	A4	Verdict
Solar-Pro4 (22B)	100	+0.241	+0.224	+0.017	+0.017	0.0130	<0.0001	✓	✓	✓	✓	**Pass**
Mistral-Small (24B)	100	+0.242	+0.217	+0.020	+0.025	0.0007	<0.0001	✓	✓	✓	✓	**Pass**
Llama-3.1 (8B)	100	+0.202	+0.200	+0.023	+0.002	0.3968	<0.0001	✓	✓	✓	×	Fail
Llama-3.2 (1B)	10	+0.019	-0.002	+0.004	+0.021	0.5332	0.5957	×	✓	×	×	Fail

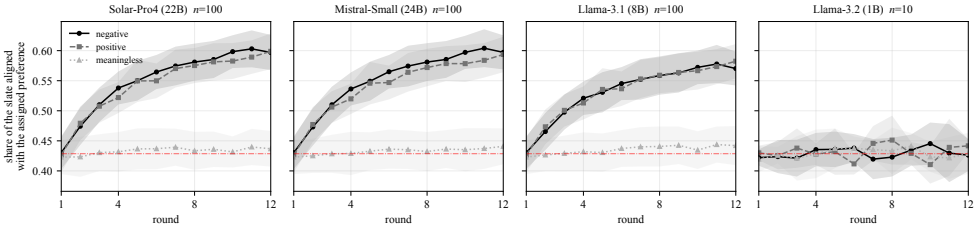


Fig. 2 Round-by-round trajectory of the share of the slate occupied by items aligned with the assigned preference.

matched representation of Section 4.3 the spread at the same adherence falls to +0.013 and stays at +0.010 even at adherence 1.00, whereas under the single bucket it remains at +0.232 or above however far adherence is raised.

What this difference does to the verdict is not small. Run under the single-bucket representation, Solar-Pro4 gives an amplification of +0.322 for the meaningless preference, above both +0.232 and +0.176 for the affective preferences, so that A3 fails. Adherence to the meaningless preference is 0.579 in that run, the lowest of the three conditions, and a result in which the condition with the weakest adherence is amplified most strongly points by itself to a defect on the side of the environment. The test therefore detects, without a generative model, an implementation defect capable of reversing the verdict.

5.3 The audit finding is reproduced by the two strongest models

Table 4 reports the four models evaluated in the repaired environment. The round-by-round trajectories are shown in Figure 2, in which the red dash-dotted line marks the initial share 30/70 of aligned items. For A1, amplification occurs in the three models other than Llama-3.2 1B. The normalised reinforcement under the negative preference is +0.242 for Mistral-Small 24B, +0.241 for Solar-Pro4 and +0.202 for Llama-3.1 8B. Llama-3.2 1B gives +0.019 under the negative preference and −0.002 under the positive one, so amplification does not occur at all and the model fails already at A1.

For A2, the two strongest models reproduce the finding of the audit. The difference is +0.025 with $p = 0.0007$ for Mistral-Small 24B and +0.017 with $p = 0.0130$ for Solar-Pro4, both significant. For these two the 95 per cent confidence interval of

Table 5 Adherence (Adh.) to the assigned preference and concentration (Conc.) within the aligned set, where ctrl. is the meaningless preference. Unparsed counts the calls from which no choice index could be recovered.

Model	Adh. (neg.)	Adh. (pos.)	Adh. (ctrl.)	Conc. (neg.)	Conc. (pos.)	Unparsed
Solar-Pro4 (22B)	0.997	0.994	0.462	0.312	0.110	38
Mistral-Small (24B)	0.991	0.992	0.429	0.197	0.127	247
Llama-3.1 (8B)	0.907	0.926	0.452	0.249	0.064	56
Llama-3.2 (1B)	0.447	0.455	0.434	0.015	0.011	254

the difference also excludes zero, as reported in Section 5.5. A3 is satisfied by three models and A4 by the top two, so Mistral-Small 24B and Solar-Pro4 satisfy every pass criterion. Llama-3.1 8B has a difference of +0.002 with $p = 0.3968$, fails A2 and therefore fails overall, and Llama-3.2 1B fails already at A1.

This verdict depends strongly on scale, and the detail is given in Section 5.5.

5.4 The mode of failure differs across models

Adherence and concentration are reported in Table 5. Adherence for Llama-3.1 8B is 0.907 under the negative preference and 0.926 under the positive one, so the model broadly follows the preference, yet the negative and positive amplifications are nearly identical and no asymmetry arises. Adherence for Llama-3.2 1B is 0.447 and 0.455, barely different from the initial aligned share of 0.43. This model effectively ignores the assigned preference. Because the premise of the audit, that users select in line with their preference, does not hold, there is no signal for the environment to learn and no amplification occurs. The failure of the 1B model is therefore not a matter of statistical power but a qualitative one of not meeting the premise of the audit.

That the mode of failure differs across models means that the procedure yields a graded diagnosis rather than a single threshold. The top two reproduce the finding of the audit, Llama-3.1 8B follows the preference without showing an asymmetry, and Llama-3.2 1B does not follow the preference at all.

The unparsed column of Table 5 counts the calls from which no choice index could be recovered. The 254 such calls for Llama-3.2 1B are tens of times as many as for the other models, yet they are 2.9 per cent of the 8,640 calls in total and do not explain why its adherence of 0.447 is no different from the initial share. The failure of this model is not a failure of output format but a consequence of not following the preference.

Adherence to the meaningless preference lies between 0.429 and 0.462 in all four models, barely different from the initial share. An instruction based on an arbitrary textual feature is not followed, regardless of model scale. In the audit the h puppet selected deterministically according to its preference and the platform did not reinforce it, whereas in this model the agents do not follow the preference and hence no signal to reinforce exists. The outcomes agree while the mechanisms do not, a point treated in Section 6.1.

This observation alone, however, cannot distinguish whether the preference is not followed because it is meaningless or cannot be followed because the feature cannot

Table 6 The two meaningless controls side by side. The initial share of aligned items is $30/70 = 0.429$ in both. The slate share is the share of aligned items each condition was actually offered, and it is what adherence should be read against.

Model	Control	n	Normalised amplification	Slate share	Adherence
Mistral-Small (24B)	h frequency	30	+0.0225	0.432	0.435
Mistral-Small (24B)	Initial letter	30	+0.0217	0.435	0.504
Solar-Pro4 (22B)	h frequency	30	+0.0166	0.433	0.464
Solar-Pro4 (22B)	Initial letter	30	+0.0075	0.437	0.489
Llama-3.1 (8B)	h frequency	30	+0.0245	0.434	0.449
Llama-3.1 (8B)	Initial letter	30	+0.0031	0.432	0.456

be computed. The control of the audit is a letter ratio over a whole transcript, which is not a feature a reader can judge at a glance. A second meaningless control was therefore used, namely whether the initial letter of the text falls in the first half of the alphabet, a feature equally without meaning but judgeable at a glance. The initial share of aligned items is matched at 30/70.

In running this control, the initial-letter feature was added to the tally of the environment. Without it, items aligned with the initial letter would have no channel through which to be reinforced and the control would be disadvantaged by construction. That change, however, came after the main conditions and the h control were run, so the two controls did not run in strictly the same environment. Table 6 therefore also reports, as the slate share, the proportion of aligned items each condition was actually offered. The share lies between 0.432 and 0.437 and is matched across the controls, so the difference of environment did not lift adherence through the composition of the slate.

The results are shown in Table 6. Adherence to the initial letter lies between 0.456 and 0.504, exceeding the slate share it was offered by only 0.024 to 0.069. A meaningless preference is not followed even when the feature is easy to judge. Amplification is also smaller for the initial letter, falling from +0.0166 to +0.0075 for Solar-Pro4 and from +0.0245 to +0.0031 for Llama-3.1 8B. What A3 captures is therefore not a failure to compute a feature but non-adherence to an instruction without meaning.

5.5 The verdict changed three times with the number of seeds

Under the same design, the same pool and the same pass criteria, the verdict changed three times with scale. At ten seeds Mistral-Small 24B passed with $p = 0.0352$, at thirty seeds no model passed, and at one hundred seeds two pass.

Between ten and thirty seeds the estimate of the effect size itself moves. The paired difference for Solar-Pro4 is +0.0413 at ten seeds and +0.014 at thirty, a fall to one third. Between thirty and one hundred seeds, by contrast, the estimate is stable: Mistral-Small 24B moves from +0.024 to +0.025 and Solar-Pro4 from +0.014 to +0.017. That is, the pass at ten seeds was a chance excursion of an unstable estimate, and the failure at thirty seeds was a shortfall of power on top of an effect

Table 7 Stability of the direction under repeated splitting, and the equivalence verdict against the pre-declared bound on a practically meaningful effect.

Model	n	Paired diff.	Sign agreement	95% CI	$\lvert\Delta\rvert \geq 0.05$ excluded
Solar-Pro4 (22B)	100	+0.0167	0.974	[+0.0019, +0.0316]	Excluded
Mistral-Small (24B)	100	+0.0250	1.000	[+0.0096, +0.0401]	Excluded
Llama-3.1 (8B)	100	+0.0020	0.207	[-0.0134, +0.0162]	Excluded
Llama-3.2 (1B)	10	+0.0214	0.460	[-0.0311, +0.0829]	Inconclusive

Table 8 Paired difference for A2 and its one-sided p on each half of the seeds. Where a half exceeds forty seeds the p is an estimate from sampling.

Model	n	Diff. first half	Diff. second half	p first half	p second half
Solar-Pro4 (22B)	100	+0.0047	+0.0287	0.3432	0.0015
Mistral-Small (24B)	100	+0.0142	+0.0358	0.1205	0.0002
Llama-3.1 (8B)	100	-0.0102	+0.0142	0.8121	0.0757
Llama-3.2 (1B)	10	+0.0575	-0.0146	0.2500	0.7812

size that had stabilised. The required seed count computed at thirty seeds was 75 for Mistral-Small 24B, and the significance obtained at one hundred seeds is consistent with that estimate.

The stability of the verdict is reported in Table 7. The seeds were split at random into halves one thousand times and the share of splits in which the sign of the paired difference agreed was computed. For comparison, Table 8 shows the result of splitting into a first and a second half once only. Under the single split Solar-Pro4 also agrees in sign, whereas under repetition its agreement rate is 0.974, and at thirty seeds it stood at only 0.706. A single split depends on how the split is taken. At one hundred seeds the rate is 1.000 for Mistral-Small 24B and 0.974 for Solar-Pro4, far above the frozen threshold of 0.80. At thirty seeds the rates were 0.889 and 0.706, and Solar-Pro4 did not reach the threshold. Llama-3.1 8B is at 0.207, so its sign does not settle even as scale increases.

The equivalence verdict against the bound $|\Delta| \geq 0.05$ on a practically meaningful effect is reported alongside. This bound is excluded for the two models that passed as well. The asymmetry is therefore detected statistically without reaching the magnitude declared in advance to be practically meaningful. For Llama-3.1 8B the confidence interval falls inside the bound, so the pre-declared bound $|\Delta| \geq 0.05$ is excluded. That does not deny the existence of a smaller asymmetry. Figure 3 traces the paired difference and its confidence interval against the number of seeds, and shows the scale required before the estimate settles.

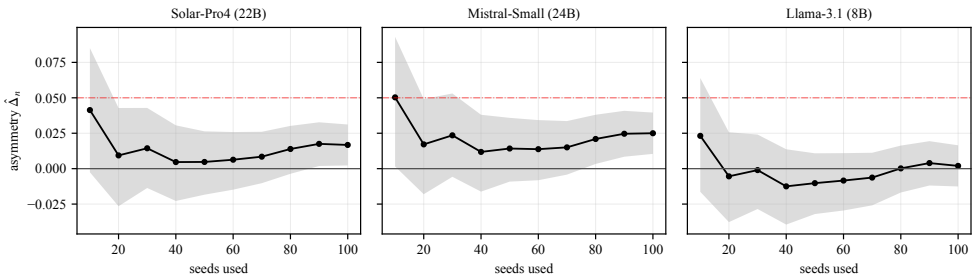


Fig. 3 Estimated asymmetry and confidence interval against the number of seeds.

Table 9 Paired difference for A2 and its one-sided p under different aggregation windows. The rightmost column measures the same quantity as a slope, which uses no window at all.

Model	n	1–3 / 10–12	1–4 / 9–12	1–2 / 11–12	Slope
Solar-Pro4 (22B)	100	+0.0167	+0.0086	+0.0116	+0.0005
Mistral-Small (24B)	100	+0.0250	+0.0161	+0.0218	+0.0010
Llama-3.1 (8B)	100	+0.0020	-0.0019	-0.0044	-0.0002
Llama-3.2 (1B)	10	+0.0214	+0.0152	-0.0135	+0.0002

5.6 How far does the verdict depend on measurement and design?

That the verdict does not depend on arbitrary design choices is checked along three axes: the aggregation window, the recommender, and the way the preference is conveyed.

5.6.1 The aggregation window

The window used to measure amplification is a design choice, and the verdict is checked not to depend on it. Table 9 places three windows alongside a measurement by the slope of the trajectory, which uses no window. Mistral-Small 24B is positive under every measurement and is significant under the slope as well, with $p = 0.018$. Solar-Pro4 is consistent in sign while its p moves from 0.013 to 0.135, so it is more sensitive to the window than Mistral-Small 24B. Llama-3.1 8B and Llama-3.2 1B change sign with the window.

At thirty seeds Solar-Pro4 also reversed sign with the window. When the scale is small, the verdict depends not only on the number of seeds but on the measurement window.

5.6.2 The recommender

The recommender of this instantiation is a simple one that samples in proportion to a tally. To check that the verdict is not specific to that mechanism, an environment based on a contextual bandit was implemented and compared under the same conditions.

Table 10 Verdict when the recommender is replaced. The pool, the slate size, the exploration rate, the prompts, the seed sequence and the pass criteria are shared.

Recommender	n	Neg.	Pos.	Meaningless	Diff.	p_{A2}
Tally (matched seeds)	30	+0.237	+0.214	+0.022	+0.0235	0.0633
Contextual bandit	30	+0.387	+0.368	-0.029	+0.0190	0.0166
Tally (all seeds)	100	+0.242	+0.217	+0.020	+0.0250	0.0007

Table 11 Verdict when the preference is conveyed differently. The initial share of aligned items is 0.429.

Conveyance	Neg.	Pos.	Meaningless	Diff.	p_{A2}	Adherence
Explicit	+0.2371	+0.2136	+0.0225	+0.0235	0.0633	0.991
Persona	+0.2115	+0.1673	+0.0098	+0.0443	0.0021	0.993
Behaviour history	+0.0009	+0.0295	+0.0050	-0.0286	0.9408	0.419

The pool, the slate size, the exploration rate, the prompts, the seed sequence and the pass criteria are shared, and what differs is only the rule that orders the items.

The results are shown in Table 10. Compared on the same thirty seeds, the paired difference is +0.0235 under the tally and +0.0190 under the bandit, so the asymmetry is preserved. Significance is in fact higher under the bandit. The level of amplification, by contrast, differs considerably: the normalised reinforcement under the negative preference rises from +0.237 to +0.387, because the bandit learns faster.

Across the two mechanisms that were tried, therefore, the level of amplification changes with the mechanism while the direction of the asymmetry does not. Since only two were compared, it cannot be said that this holds for every recommender. For the meaningless preference the bandit turns negative, at -0.029, so this environment does not learn the arbitrary textual feature and moves away from it. The control of A3 functions under a changed environment as well.

5.6.3 The way the preference is conveyed

The sock puppets of the audit are given their preference as a program. A real user is not told a preference but carries it over in the form of a history of what has been opened. To check whether this difference matters to the verdict, the same preference was conveyed in three forms and compared on Mistral-Small 24B. The numbers in this subsection are based on the thirty seeds on which all three forms were run, so the paired difference of the explicit condition does not match the value at one hundred seeds. The explicit form states the preference as an instruction and is the form used in the runs so far. The persona form describes the preference as a disposition. The behaviour-history form states no preference at all and shows only the posts that the agent opened in the past.

Table 12 Kullback–Leibler divergence from the frozen pool when a distribution-level criterion is applied to the same runs.

Model	n	Negative pref.	Positive pref.	Meaningless pref.
Solar-Pro4 (22B)	100	1.119	0.844	0.174
Mistral-Small (24B)	100	0.975	0.898	0.299
Llama-3.1 (8B)	100	0.783	0.617	0.079
Llama-3.2 (1B)	10	0.047	0.048	0.037

The results are shown in Table 11. Conveyed as a persona, the asymmetry in fact strengthens. The paired difference moves from +0.0235 to +0.0443 and p from 0.0633 to 0.0021. In the breakdown the negative amplification barely changes while the positive amplification falls from +0.2136 to +0.1673. When being drawn to calm and cheerful things is written as a disposition, adherence to the positive preference weakens in relative terms.

Conveyed as a behaviour history, the reproduction is lost entirely. Amplification under the negative preference is +0.0009 and the asymmetry reverses in sign to −0.0286. The cause lies in adherence. It is 0.99 under both the explicit and the persona forms and 0.419 under the behaviour history, barely different from the initial aligned share of 0.429. The agents, that is, do not carry a preference over from a history of what they themselves have opened.

This result shows what the reproduction in this study depends on. The situation the audit modelled was one in which a user keeps selecting in line with their own preference. This instantiation could reproduce it because the preference was supplied as language, and not because a preference was reconstructed from a history.

5.7 Where did the agreement come from?

5.7.1 A distribution-level criterion does not separate the same runs

To see what audit replication adds to distribution-level validation, a distribution-level criterion was applied to the same runs. The category distribution of the items selected in the late rounds was compared against the distribution a platform without preferences would deliver, namely the distribution of the frozen pool itself, and the Kullback–Leibler divergence was computed. The results are shown in Table 12.

The numbers below are based on the same one hundred seeds as the main results. The three models other than Llama-3.2 1B depart substantially from the pool in every preference condition. Read on the distribution alone, all three would be taken to change their delivery in response to the preference. Under the verdict of audit replication, however, the three are not the same. The top two reproduce the negative and positive asymmetry at conventional significance, while Llama-3.1 8B shows no asymmetry even at one hundred seeds. A distribution-level criterion does not separate this difference. What it detects is whether the system responds to the preference, and not whether the direction of the asymmetry reported by the audit is reproduced.

Table 13 The attribution ladder. Each rung is a random chooser that adds one constraint to the rung below it, and the gap to the observed value is an upper bound on the residual that the agent-side summaries do not explain.

Model	n	Observed	Null 0	Null 1	Null 2	Null 3	Null 4
Solar-Pro4 (22B)	100	+0.0167	+0.0013	+0.0111	+0.0086	+0.0075	+0.0066
Mistral-Small (24B)	100	+0.0250	+0.0013	+0.0088	+0.0059	+0.0087	+0.0074
Llama-3.1 (8B)	100	+0.0020	+0.0013	-0.0018	-0.0050	-0.0027	-0.0033
Llama-3.2 (1B)	10	+0.0214	+0.0070	+0.0020	-0.0114	-0.0175	-0.0065

5.7.2 How much of the asymmetry the choice distribution explains

To separate whether the asymmetry of A2 is due to the action of the environment or to the choices of the agents themselves, the attribution ladder is shown in Table 13. Each rung is a random chooser that adds one constraint to the rung below, so that the higher the rung the more finely it reproduces the behaviour of the agents. All numbers below are based on one hundred seeds.

Under uniform selection almost no asymmetry arises. It is +0.0013 for both Solar-Pro4 and Mistral-Small 24B, so no difference between negative and positive appears as long as selection is made without a preference. At the rung that matches adherence alone this jumps. For Solar-Pro4 it is +0.0111, two thirds of the observed +0.0167, and for Mistral-Small 24B +0.0088, one third of the observed +0.0250. A substantial part of the asymmetry, that is, arises from the single quantity of how far the preference is followed.

The rungs that add per-agent constraints in fact move away from the observed value. Reproducing each agent's distribution exactly fixes the variation between agents and thereby weakens the amplification of the population, which is how this is interpreted. The rungs therefore do not accumulate monotonically, and the rung closest to the observed value is the one that matches adherence alone.

Because the ladder does not accumulate monotonically, the residual is defined not as the distance from the highest rung but as the distance from the rung that best approximates the observed value. Writing Δ_{obs} for the observed asymmetry and $\Delta_{\text{null}}^{(j)}$ for the baseline at rung j,

$$j^* = \arg\min_j \left| \Delta_{\text{obs}} - \Delta_{\text{null}}^{(j)} \right|, \qquad \Delta_{\text{res}} = \Delta_{\text{obs}} - \Delta_{\text{null}}^{(j^*)} \tag{2}$$

This is not the contribution of the environment itself but an upper bound on the residual that the agent-side summaries do not explain. In the runs of this paper every rung fell below the observed value, so the closest rung coincides with the rung of maximum value. When a baseline exceeds the observed value the two differ, and Equation (2) is used in that case as well.

Under this definition the residual is +0.0057 for Solar-Pro4 and +0.0162 for Mistral-Small 24B.

Figure 4 places the observed asymmetry beside the baseline and shows that the baseline is close to the observed value for Solar-Pro4 while a larger residual remains for Mistral-Small 24B. It can therefore be said that what produces A2 is largely

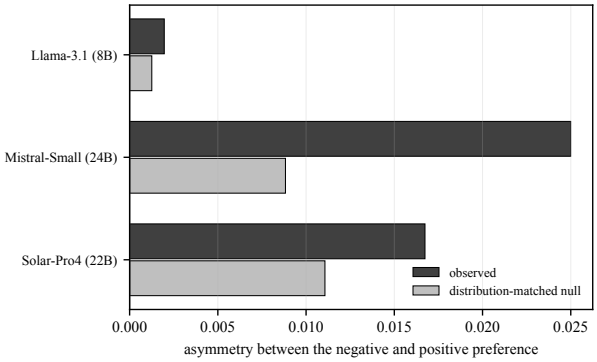


Fig. 4 The observed asymmetry and the distribution-matched null baseline.

Table 14 Normalised amplification regressed on adherence and concentration. The coefficient of determination is the share explained by the agent-side quantities alone.

Model	n	Observations	R^2 adherence	R^2 adherence+concentration	$\beta_{\text{concentration}}$
Solar-Pro4 (22B)	100	200	0.000	0.031	+0.073
Mistral-Small (24B)	100	200	0.000	0.038	+0.156
Llama-3.1 (8B)	100	200	0.012	0.014	+0.019
Llama-3.2 (1B)	10	20	0.100	0.102	-0.150

explicable by the choices of the agents for Solar-Pro4, whereas for Mistral-Small 24B the unexplained part is the larger one. Whether that residual is due to the action of the environment or to chance cannot be decided from the data of this study. For Llama-3.1 8B the observed asymmetry is itself +0.0020, so there is no asymmetry to explain.

The skew of the choices is also worth noting. For Solar-Pro4 the concentration is 0.312 under the negative preference and 0.110 under the positive one, so selections under the negative preference are strongly skewed towards anger. For Mistral-Small 24B they are 0.197 and 0.127, a much smaller difference in skew. That the magnitude of the asymmetry is comparable although the difference in concentration differs greatly shows that concentration alone cannot be said to cause the asymmetry. That anger propagates more strongly than other emotions has been reported on real data (Fan et al. 2014), and the behaviour of this model is consistent with that direction.

Table 14 examines the same question as the ladder by regression. Regressing the per-seed normalised reinforcement on adherence and concentration, the coefficient of determination is below 0.04 for every model. Adherence sets the level across models but does not explain the variation across seeds within a model. The spread of amplification derives from the composition of the slate and the chance of selection, and is not captured by agent-side summaries.

5.8 The procedure does not manufacture positive results

On top of the validated environment, a regime for which no audit exists was placed and run under the same discipline. The object is a feed that optimises the emotion of the recipient, with the learned signal replaced by the emotional change measured from selections. As comparison conditions, random delivery, topic matching, delivery learned from selections, and random delivery with the delivery volume matched exactly were placed.

The delivery check of element 7 passed. The mean feed valence of the affect-optimising condition is +0.168, which exceeds the +0.016 of random delivery by 0.152. The treatment was in fact delivered.

The delivery check passes against random delivery, yet the mean feed valence is +0.182 under topic matching and +0.160 under delivery learned from selections, so the affect-optimising condition does not exceed them. Setting emotion as the objective did not deliver positive content specifically more often.

Of the four predictions registered in advance, one held. The rank correlation between delivery volume and the emotional state of the recipient is −0.320, with $p = 0.0022$ under a two-sided test, far from the calibrated-null baseline of +0.019. The sign is negative, so agents in a higher emotional state receive fewer items. The standard deviation across seeds is 0.308, however, comparable to the mean, and the value on individual seeds ranges from −0.029 to −0.753. The direction of the effect agrees on all ten seeds while its magnitude is not stable.

Moreover, the same correlation is −0.188 under the condition of random delivery with the delivery volume matched exactly. Because the series of delivery volumes is itself derived from the dynamics of the affect-optimising condition, merely pouring it into another society reproduces 59% of the observed correlation. Applying the attribution analysis proposed in this paper to this application, what can be called specific to affective selection is only −0.132 of the −0.320. The difference in the accuracy of reading the emotions of others is −0.001, the difference in the rate of remaining in the lower band is −0.001, and the difference in the response to a common shock is −0.012, all practically zero.

For the one prediction that held, a change of criterion must be disclosed. For the correlation between delivery volume and emotional state, the direction initially expected was that agents in a lower emotional state receive less, and the criterion was $\rho \geq +0.30$. Because the null-agent test showed that the correlation generated by the admissibility mechanism runs in the opposite direction, the criterion was changed to the two-sided $|\rho| \geq 0.30$ and the sign was made something to be reported as a result. The observed sign is opposite to the direction initially expected. Since this study cannot provide a means by which a third party could independently verify the chronology of that change, it is stated as a record of what happened and not as a claim of priority. The one prediction that held in this subsection is therefore treated as an exploratory diagnostic and not as a confirmatory finding. It is not used as evidence for the validity of the procedure in the discussion that follows.

Table 15 summarises the arm. The treatment is delivered, yet the mean feed valence of the affect-optimising condition does not exceed that of topic matching, and the volume-matched random condition alone recreates 59% of the rank correlation. That

Table 15 Application to a regime for which no audit finding exists. The mean feed valence shows whether the treatment was delivered, and the rank correlation is the statistic of the prediction registered in advance.

Condition	n	Volume	Mean feed valence	Rank correlation
Random delivery	10	6.00	+0.016	–
Topic matching	10	6.00	+0.182	–
Learned from selections	10	6.00	+0.160	–
Affect optimising	10	5.78	+0.168	-0.320
Volume-matched random	10	5.78	+0.104	-0.188

three predictions failed while the treatment was delivered shows that the procedure can be applied regardless of whether a prediction holds. For a validation procedure to be useful, it must be confirmed that it does not manufacture positive results.

6 Discussion of Results and Implications

6.1 Interventional agreement does not establish mechanistic agreement

The principal finding of this study is that, once the finding of an audit has been reproduced, how far that reproduction reaches towards mechanism can be given as a number.

The two strongest models satisfy A1 through A4, so they agree at the level of the intervention. The attribution analysis, however, shows that a substantial part of the observed asymmetry is reproduced by a baseline that is given only the observed adherence and choice distribution. The share so reproduced is two thirds for Solar-Pro4 and one third for Mistral-Small 24B, and thus differs across models. In either case the agreement cannot be attributed uniquely to the amplification mechanism of the environment. A similar qualification is needed for A3. Adherence to the meaningless preference is barely different from the initial share in all four models, so in the audit the preference was obeyed and the platform did not pick it up, whereas in this model the agents do not obey the preference and hence no signal to pick up exists. The outcomes agree while the mechanisms do not.

A qualification is needed for the magnitude of the reproduction as well. Even for the two models that passed, the confidence interval of the paired difference excludes the pre-declared practical bound $|\Delta| \geq 0.05$. The finding of the audit is reproduced statistically, while its magnitude does not reach the level this study decided in advance to regard as practically meaningful. Whether a reproduction holds and whether its effect matters in practice are separate questions.

This decomposition is a result and not a limitation. That distributional agreement does not imply mechanistic agreement has already been pointed out (Zhao et al. 2026; Li and Ji 2026), and what this study adds is the further step that interventional agreement does not establish mechanistic agreement either. The size of that gap, moreover, can be measured without any additional generative call.

6.2 What decided the verdict

Three things decided the verdict: diagnostics that use no generative model, the size of the sample, and the way the preference is conveyed.

Without the null-agent test of element 3, the failure of A3 under the single-bucket representation might have been reported as the finding that LLM agents respond even to a meaningless preference, whereas the cause lay in the feature space of the environment, as the difference of +0.271 in amplification at identical adherence shows. Elements 4 and 7 acted in the same way. In the application to the affect-optimising regime the initial implementation learned per individual, and against a measured response separation of +0.063 the standard deviation per observation was 0.19, so over twelve rounds the estimate was buried under the noise. The treatment then acted not on the selection of content but as a nearly uniform reduction of delivery volume, and without the delivery check that run would have been reported as the negative result that an affect-optimising feed has no macro effect. Generalised, the lesson is to measure the signal-to-noise ratio of the mechanism to be learned before designing the learner, because the responses of LLM agents are not necessarily large compared with the effect sizes reported for humans. Delivery and reception of a treatment need to be distinguished in the same breath. Llama-3.2 1B was given the same preference-stating prompt as the other models, so delivery holds and what does not hold is reception. Void is a category used only when delivery fails, and a treatment that is delivered but not received is a failure.

Under the same design, the same pool and the same pass criteria, the verdict changed three times with scale, and the quality of the change differs between the intervals. Between ten and thirty seeds the estimate of the effect size itself moves, the paired difference for Solar-Pro4 falling from +0.0413 to +0.014. The paired difference over the twenty added seeds is +0.0009 and the standard deviation barely changes, so the spread did not increase and the first ten seeds had been a chance excursion upward. Between thirty and one hundred seeds the estimate is stable and what changed is power. This experience shows the limitation of element 6: computing the required number of seeds depends on an estimate of the effect size, and as long as that estimate comes from the same small number of seeds the computation is circular. Element 6 was therefore strengthened to demand a split replication, and the verdict of a model whose halves do not agree must not be treated as a stable finding. Comparing against the effect sizes the audit reported places the scale of this study in perspective, since the audit reports the difference between the negative and the positive preference as +0.53 in the News domain and +0.20 in the Fitness domain, yet it reverses in sign to −0.23 in the Random domain. The asymmetry the audit reported is therefore domain-dependent, and it is entirely possible for it to be small in the context to which this instantiation corresponds. The verdict does not change with the treatment of multiple comparisons, since the two strongest models remain below the significance level under the Bonferroni correction.

The third factor is the way the preference is conveyed. When the same preference is supplied as language, adherence reaches 0.99, whereas when only the posts the agent itself opened in the past are shown, adherence is 0.419, no different from the initial aligned share of 0.429. An LLM agent does not by default carry a preference over from

its own behavioural history. The sock puppets of the audit hold their preference as a program throughout, whereas a real user sustains it as behaviour. This instantiation could reproduce the finding because the preference was supplied as language, and not because a preference was reconstructed from a history. More broadly, a simulation designed to hold a behavioural history as state needs to confirm that the state actually governs subsequent behaviour.

6.3 Theoretical and practical implications

Theoretically, these results separate validity into three levels rather than two. Agreement of distributions is the weakest, agreement of the response to an intervention is stronger, and agreement of mechanism is stronger still, and the present results show that the second does not deliver the third even when the second is established with a pre-declared direction, a transplanted control and a corrected significance level. The literature that moved from distributional comparison (Tomašević et al. 2026) to interventional criteria (Cross et al. 2025; Cui et al. 2024) has treated the second level as the destination; this study measures the remaining distance to the third and finds it large enough that two models which pass every criterion owe one third and two thirds of the reproduced effect to their own observable behaviour. This also connects the validation of simulated users to the older concern of evaluation in information access, where a measure is trusted only once the conditions under which it stops discriminating are known (Voorhees 2000; Valcarce et al. 2020). The distribution-matched null is such a condition, made explicit and quantitative. Compared with the user-simulation framework that requires predictive validity against real users (Balog and Zhai 2024), the procedure proposed here relaxes the data requirement from real behaviour to a published directional finding, which is what makes it applicable to a platform no researcher can run an experiment on.

Practically, the results give anyone who evaluates a recommender through simulated users four things to do before reading the output as evidence. First, run the environment with the generative model replaced by a fixed-probability chooser, because a feature space that favours one condition by construction produced a difference of +0.271 here at identical adherence and would otherwise have been reported as a property of the agents. Second, measure the separation and the noise of the signal the environment is meant to learn, since a treatment can be delivered and still leave no learnable trace. Third, report a distribution-matched null beside every positive result, which costs no additional generative call and bounds how much of the result the agents' own behaviour already explains. Fourth, decide the number of runs by split replication rather than by a power calculation that reuses the effect size estimated from the same small sample, since the verdict here changed three times between ten and one hundred runs while the effect size did not. For work that proposes agent simulation as a substitute for offline or online testing (Wang et al. 2025; Bougie et al. 2026), these four steps are the difference between a simulation that agrees with a platform and a simulation that can be said to agree for the platform's reason.

6.4 Limitations

First, the object of reproduction is a single audit. Claiming generality requires applying the procedure to audits with a different target system, intervention and observable. The application here is in particular confined to the relation between an individual and a platform, and has not been applied to an audit that involves social dynamics.

Second, it cannot be concluded that no asymmetry exists for Llama-3.1 8B. Because the confidence interval at one hundred seeds falls inside the practical bound, it can only be said that there is no asymmetry of a magnitude detectable at this scale. Llama-3.2 1B remains at ten seeds, and no argument about power is made for that model.

Third, the reproduction depends on the way the preference is conveyed, so it holds only for the part of the audit's setting in which the preference is stated explicitly. No model passed the calibration either, so the application arm was run only on models whose calibration remained undetermined.

Fourth, the content pool was created by a generative model; the categories are known by construction, yet the pool may reflect the vocabulary and the style of a particular model. This paper further takes the emotion categories attached to the texts as given. There is a fundamental criticism of inferring emotion from expression at all (Barrett et al. 2019), and the framework here carries over the operationalisation of emotion that the audit used rather than validating it independently.

7 Conclusion

Evaluating an information access system through simulated users requires that the simulation be valid, and the criterion in use demands either agreement with the behaviour of real users or agreement of distributions with a target population. This study constructed a procedure that instead reproduces the directional findings of a published algorithm audit in silico, which removes the requirement for data on real users, and measured by attribution analysis how far such a reproduction reaches towards mechanism. The procedure consists of fixing the direction in advance, transplanting the audit's own control condition, pre-expenditure tests that use no generative model, the attribution analysis, a check of design sensitivity and a check of delivery, and its verdict takes the four values pass, failure, inconclusive and void.

Applied to four generative models against an audit of affective recommendation, the two strongest models reproduced the finding once one hundred seeds were taken. A baseline given only the observed adherence and choice distribution, however, recreates two thirds of the asymmetry for one model and one third for the other, so agreement at the level of the intervention holds while agreement at the level of mechanism is not established. The pre-expenditure tests remade the verdict, the verdict changed three times between ten and one hundred runs while the effect size did not, and conveying the preference through behavioural history rather than language removed the reproduction entirely.

Three tasks remain. First, to extend the object of reproduction to several audits and examine whether the procedure works for a different target system, intervention and observable. Second, to use several prompt styles and content pools and confirm

that the verdict does not depend on them. Third, to ask whether agreement at the level of mechanism appears once the fidelity of the environment is raised. What this study has shown is that at the present level interventional agreement does not reach mechanistic agreement, and how far the environment must be refined before it does remains open.

Acknowledgements. None.

Declarations

Funding. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Competing interests. The author has no competing interests to declare that are relevant to the content of this article.

Ethics approval and consent to participate. Not applicable. The study involves no human participants and no data on human participants.

Consent for publication. Not applicable.

Data availability. The frozen content pool, the pass criteria and every stored run reported in this paper are available at https://github.com/keito-git/audit-replication-nulls.

Code availability. The code that runs the protocol and generates every table and figure reported in this paper is available at the same repository, https://github.com/keito-git/audit-replication-nulls.

Author contribution. The author conceived the study, designed the protocol, implemented the software, ran the experiments, analysed the results, prepared the figures and wrote the manuscript.

Use of generative artificial intelligence. During the preparation of this work the author used generative artificial intelligence tools to assist with coding, and to proofread and translate the text. After using these tools the author reviewed and edited the content as needed and takes full responsibility for the content of the article. The language models evaluated in the study are described in Section 5.1 and are objects of study rather than tools of authorship.

References

Barrett, L.F., Adolphs, R., Marsella, S., Martinez, A.M., Pollak, S.D.: Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements. Psychological Science in the Public Interest **20**(1), 1–68 (2019) https://doi.org/10.1177/1529100619832930

Bandy, J.: Problematic machine behavior: A systematic literature review of algorithm audits. Proceedings of the ACM on Human-Computer Interaction **5**(74), 1–34 (2021) https://doi.org/10.1145/3449148

Bougie, N., Marconi, G.M., Ye, X., Watanabe, N.: Beyond offline A/B testing: Context-aware agent simulation for recommender system evaluation. arXiv (2026) https://doi.org/10.48550/arXiv.2604.09549

Bonabeau, E.: Agent-based modeling: Methods and techniques for simulating human systems. Proceedings of the National Academy of Sciences **99**(suppl_3), 7280–7287 (2002) https://doi.org/10.1073/pnas.082080899

Bougie, N., Watanabe, N.: SimUSER: Simulating user behavior with large language models for recommender system evaluation. arXiv (2025) https://doi.org/10.48550/arXiv.2504.12722

Balog, K., Zhai, C.: User simulation for evaluating information access systems. Foundations and Trends in Information Retrieval **18**(1-2), 1–261 (2024) https://doi.org/10.1561/1500000098

Balog, K., Zhai, C.: User simulation in the era of generative AI: User modeling, synthetic data generation, and system evaluation. arXiv (2025) https://doi.org/10.48550/arXiv.2501.04410

Chuang, Y.-S., Goyal, A., Harlalka, N., Suresh, S., Hawkins, R., Yang, S., Shah, D., Hu, J., Rogers, T.: Simulating opinion dynamics with networks of LLM-based agents. In: Findings of the Association for Computational Linguistics: NAACL 2024, pp. 3326–3346 (2024). https://doi.org/10.18653/v1/2024.findings-naacl.211

Cross, L., Haber, N., Yamins, D.L.K.: Validating generative agent-based models of social norm enforcement: From replication to novel predictions. arXiv (2025) https://doi.org/10.48550/arXiv.2507.22049

Cui, Z., Li, N., Zhou, H.: Can large language models replace human subjects? a large-scale replication of scenario-based experiments in psychology and management. arXiv (2024) https://doi.org/10.48550/arXiv.2409.00128

Durupinar, F.: How affect propagates among LLM agents: Emergent emotional contagion in crowd simulation. arXiv (2026) https://doi.org/10.48550/arXiv.2607.25140

Fan, R., Zhao, J., Chen, Y., Xu, K.: Anger is more influential than joy: Sentiment correlation in Weibo. PLoS ONE **9**(10), 110184 (2014) https://doi.org/10.1371/journal.pone.0110184

Grimm, V., Revilla, E., Berger, U., Jeltsch, F., Mooij, W.M., Railsback, S.F., Thulke, H.-H., Weiner, J., Wiegand, T., DeAngelis, D.L.: Pattern-oriented modeling of agent-based complex systems: Lessons from ecology. Science **310**(5750), 987–991 (2005) https://doi.org/10.1126/science.1116681

Hammond, R.A., Barkin, S.: Making evidence go further: Advancing synergy between

agent-based modeling and randomized control trials. Proceedings of the National Academy of Sciences **121**(21), 2314993121 (2024) https://doi.org/10.1073/pnas.2314993121

Hullman, J., Broska, D., Sun, H., Shaw, A.: This human study did not involve human subjects: Validating LLM simulations as behavioral evidence. arXiv (2026) https://doi.org/10.48550/arXiv.2602.15785

Habib, H., Nithyanand, R.: YouTube recommendations reinforce negative emotions: Auditing algorithmic bias with emotionally-agentic sock puppets. arXiv (2025) https://doi.org/10.48550/arXiv.2501.15048

Inoshita, K.: Bridging the silos in affective AI: A critical perspective from data to society. AI & Society (2026) https://doi.org/10.1007/s00146-026-03324-y

Inoshita, K.: Who determines the meaning of an emotion? Affective sovereignty as an epistemic consequence of measurement limits. arXiv (2026) https://doi.org/10.48550/arXiv.2606.31442

Inoshita, K., Zhou, X., Kawai, A., Yada, K.: LLMs capture emotion labels, not emotion uncertainty: Distributional analysis and calibration of human–LLM judgment gaps. arXiv (2026) https://doi.org/10.48550/arXiv.2604.27345

Li, Z., Ji, F.: Statistical realism is not evidence that LLMs can estimate treatment effects in social science experiments. arXiv (2026) https://doi.org/10.48550/arXiv.2604.02458

Larooij, M., Törnberg, P.: Do large language models solve the problems of agent-based modeling? a critical review of generative social simulations. arXiv (2025) https://doi.org/10.48550/arXiv.2504.03274

Luo, S., Yao, S., Zhang, D.J.: Unintended consequences of recommender system interventions: Evidence from a field experiment. arXiv (2026) https://doi.org/10.48550/arXiv.2606.08265

Mollahosseini, A., Hasani, B., Mahoor, M.H.: AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild. IEEE Transactions on Affective Computing **10**(1), 18–31 (2019) https://doi.org/10.1109/TAFFC.2017.2740923

Ojewale, V., Steed, R., Vecchione, B., Birhane, A., Raji, I.D.: Towards AI accountability infrastructure: Gaps and opportunities in AI audit tooling. In: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, pp. 1–29 (2025). https://doi.org/10.1145/3706598.3713301

Pandey, A., Jajoo, G.: Reason-mediated behavioral models for auditing LLM social simulators. arXiv (2026) https://doi.org/10.48550/arXiv.2607.24649

Park, J.S., O'Brien, J., Cai, C.J., Morris, M.R., Liang, P., Bernstein, M.S.: Generative agents: Interactive simulacra of human behavior. In: Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, pp. 1–22 (2023). https://doi.org/10.1145/3586183.3606763

Piao, J., Yan, Y., Zhang, J., Li, N., Yan, J., Lan, X., Lu, Z., Zheng, Z., Wang, J.Y., Zhou, D., Gao, C., Xu, F., Zhang, F., Rong, K., Su, J., Li, Y.: AgentSociety: Large-scale simulation of LLM-driven generative agents advances understanding of human behaviors and society. arXiv (2025) https://doi.org/10.48550/arXiv.2502.08691

Qiang, Q.: Emotion diffusion in real and simulated social graphs: Structural limits of LLM-based social simulation. arXiv (2025) https://doi.org/10.48550/arXiv.2512.21138

Russell, J.A.: A Circumplex Model of Affect. Journal of Personality and Social Psychology **39**(6), 1161–1178 (1980) https://doi.org/10.1037/h0077714

Sargent, R.G.: Verification and validation of simulation models. Journal of Simulation **7**(1), 12–24 (2013) https://doi.org/10.1057/jos.2012.20

Stray, J., Vendrov, I., Nixon, J., Adler, S., Hadfield-Menell, D.: What are you optimizing for? aligning recommender systems with human values. ICML 2020 Participatory Approaches to Machine Learning workshop (2020) https://doi.org/10.48550/arXiv.2107.10939

Sukiennik, N., Wang, H., Zeng, Z., Gao, C., Li, Y.: Simulating filter bubble on short-video recommender system with large language model agents. arXiv (2025) https://doi.org/10.48550/arXiv.2504.08742

Tomašević, A., Cvetković, D., Major, S., Maletić, S., Anđelković, M., Vranić, A., Stupovski, B., Vudragović, D., Bogojević, A., Mitrović Dankulov, M.: Towards operational validation of LLM-agent social simulations: a replicated study of a Reddit-like technology forum. EPJ Data Science **15**, 72 (2026) https://doi.org/10.1140/epjds/s13688-026-00674-x

Valcarce, D., Bellogín, A., Parapar, J., Castells, P.: Assessing ranking metrics in top-N recommendation. Information Retrieval Journal **23**(4), 411–448 (2020) https://doi.org/10.1007/s10791-020-09377-x

Voorhees, E.M.: Variations in relevance judgments and the measurement of retrieval effectiveness. Information Processing and Management **36**(5), 697–716 (2000) https://doi.org/10.1016/S0306-4573(00)00010-8

Windrum, P., Fagiolo, G., Moneta, A.: Empirical validation of agent-based models: Alternatives and prospects. Journal of Artificial Societies and Social Simulation **10**(2), 8 (2007)

Wang, L., Zhang, J., Yang, H., Chen, Z.-Y., Tang, J., Zhang, Z., Chen, X., Lin, Y., Sun, H., Song, R., Zhao, X., Xu, J., Dou, Z., Wang, J., Wen, J.-R.: User behavior simulation with large language model-based agents. ACM Transactions on Information Systems **43**(2), 1–37 (2025) https://doi.org/10.1145/3708985

Ye, J., Cao, L., Chen, D., Ferrara, E.: Stop drawing scientific claims from LLM social simulations without robustness audits. arXiv (2026) https://doi.org/10.48550/arXiv.2605.18890

Yeykelis, L., Pichai, K., Cummings, J.J., Reeves, B.: Using large language models to create AI personas for replication, generalization and prediction of media effects: An empirical test of 133 published experimental research findings. arXiv (2024) https://doi.org/10.48550/arXiv.2408.16073

Yang, Z., Zhang, Z., Zheng, Z., Jiang, Y., Gan, Z., Wang, Z., Ling, Z., Chen, J., Ma, M., Dong, B., Gupta, P., Hu, S., Yin, Z., Li, G., Jia, X., Wang, L., Ghanem, B., Lu, H., Lu, C., Ouyang, W., Qiao, Y., Torr, P., Shao, J.: OASIS: Open agent social interaction simulations with one million agents. arXiv (2025) https://doi.org/10.48550/arXiv.2411.11581

Zhao, P., Pham, D.H., Vincent, N.: Mechanism plausibility in generative agent-based modeling. In: The 2026 ACM Conference on Fairness, Accountability, and Transparenc, pp. 5915–5940 (2026). https://doi.org/10.1145/3805689.3812388