WHO WAKES UP? EVENT-CONDITIONED ACTIVATION OF AGENTS IN SOCIAL SIMULATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Large-scale social simulation with LLM agents cannot afford to activate every agent for every event, and existing systems resolve this with event-blind, time-driven scheduling that ignores what the event is about. We argue that activation is a behavioral mechanism and study it as such — *event-conditioned agent activation*: given an event, which agents should wake up? We formalize activation as ranking over the full agent pool, map eight policy families spanning heuristic scheduling, dense retrieval, collaborative filtering, and discriminative fine-tuning, and evaluate them under a deployment-faithful protocol: all 6,566 agents scored on 2,241 real events, multi-seed, with paired significance tests. A single lens organizes everything we find: the gap between the negative-sampling distribution a policy is trained or evaluated on and the full-pool distribution it faces at deployment. Balanced 1:1 training collapses under full-pool ranking, and moving toward the deployment ratio only makes things worse: at 1:500 — still well below the true deployment ratio of $\sim$1:1,750 — two of three seeds fail and the survivor scores 2.9% P@1, the worst configuration we tested. Negatives with *verified event exposure* — users who liked but never replied — beat same-seed random negatives at the widest tested ratios and tame run-to-run variance; whether the gain comes from exposure realism or negative hardness, our design cannot separate. A frozen dense retriever ranks as well as a trained classifier, but its scores have no usable operating point: a fixed 0.5 threshold selects 575$\times$ the true number of responders, versus 2.1$\times$ for the trained classifier. And held-out metrics miss all of it — one in six training runs collapses silently. The practical upshot: train one to two orders of magnitude below the deployment ratio, mine negatives from verified exposure, let retrieval recall and a discriminative model decide, and evaluate on the full pool across seeds. Activation is a behavioral mechanism, not a scheduling detail — and what it is trained and measured on matters as much as the model.

1 INTRODUCTION

Large language model (LLM) agents have made it possible to simulate social systems at unprecedented scale: recent frameworks run hundreds of thousands to millions of agents that post, reply, and redistribute content in replica social environments (Park et al., 2023; Gao et al., 2023; Yang et al., 2024). Yet scale has sharpened a question that these systems answer only implicitly: *when an event happens, which agents should wake up and respond?* Activating every agent for every event is computationally infeasible at scale. Existing systems therefore resolve activation with time-driven scheduling — hourly activity coefficients, random sampling from the agent pool, or fixed activity levels assigned at initialization (Yang et al., 2024). In such designs, activation is a background scheduling detail; functionally, however, it is the mechanism that selects the *participants* of every simulated interaction — the support set on which all downstream generation unfolds.

We argue that agent activation should be studied as a first-class, event-conditioned *behavioral* mechanism. Real participation is not a random time process: who responds to an event depends on what the event is about, who the user is, and what they have cared about before. An activation policy that ignores event semantics does not merely add noise — it systematically distorts who enters the interaction field, biasing cascades and participant composition before any content is generated.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

This paper studies *event-conditioned agent activation* as a ranking problem: given an event and a pool of thousands of agents, select the small subset who would plausibly respond. We formalize the problem (§3), map eight policy families from heuristic scheduling to LLM judgment (§4), and evaluate them under a deployment-faithful protocol — full-pool ranking over 2,241 test events and 6,566 agents from real Mastodon cascades, with cross-platform checks on Zhihu (§5) — using paired, multi-seed evaluation with exact significance testing throughout. Our contributions: the problem formulation, the eight-family design-space map, the evaluation protocol, and the four findings below.

A single lens organizes this entire design space. At deployment, an activation policy faces the full agent pool: every agent is a candidate, and the overwhelming majority are negatives — an average event draws 3.7 unique responders from 6,566 agents, an effective ratio of $\sim$1:1,750. The negative-sampling distribution a policy is trained and evaluated on is therefore a design choice in its own right, and its gap to the full-pool deployment distribution predicts — better than model capacity or policy family — what works, what fails, and what remains invisible to standard evaluation (§6, §7):

- **Negative-exposure intensity is a dial with a wide safe zone and two cliffs.** Balanced 1:1 training starves the decision boundary and collapses under full-pool ranking; the plateau at 1:5–1:70 is wide enough that exact tuning is unnecessary; and moving toward the deployment ratio only makes things worse: at 1:500, two of three seeds are lost to the all-negative attractor and the survivor scores 2.9% P@1 — the worst configuration we measured, below even balanced training — while the true deployment ratio lies $3.5\times$ further out, deep past the collapse region. The standard advice to "match the training distribution to deployment" points straight into it.
- **Where negatives come from matters more than how many.** Random negatives are drawn blind: most never saw the event at all. Replacing them with *verified-exposure* negatives (users who liked the seed post but never replied) beats same-seed random twins at 1:50 in both seeds and tames seed variance from $\sim$5pp to $\sim$1pp. The benefit is nil at low ratios and still mixed at 1:30, where one seed collapses: hard negatives pay off only once numerous enough to shape the boundary.
- **Ranking and operating points decouple.** A frozen e5 retriever matches the trained classifier's full-pool P@1 (no significant difference detected) with substantially complementary wins (an oracle union adds $+$7pp) — retrieval gets the *order* right without any task training. But its scores have no usable operating point: a fixed 0.5 threshold selects $575\times$ the true responder count, where the 1:50 classifier lands at $2.1\times$. The deployable consequence is a division of labor: retrieval for recall, a discriminative model for decisions.
- **Off-distribution evaluation hides the failures that matter.** Held-out pair classification lives off the deployment distribution, and it cannot see full-pool failure: across 36 runs we find two invisible modes — *ranking collapse* (held-out F1 normal, full-pool P@1 off a cliff) and the *all-negative attractor* at extreme ratios (F1 $= 0$ while the encoder keeps learning) — roughly one run in six. Single-run numbers swing several points between seeds; a right-tail "rebound" in our own early curve was a seed artifact that replication falsified. We therefore report every headline result multi-seed with same-seed paired tests.

These results form a design map held together by a single lens — mind the gap between the training and deployment distributions, and close it where it is cheap: train one to two orders of magnitude below the deployment ratio, mine negatives from verified exposure, pair a retriever's recall with a discriminative model's decisions, and evaluate where you deploy, over as many seeds as you can afford.

2 Related Work

LLM-agent social simulation. Generative agents that maintain memory, plan, and interact in natural language have been scaled from tens of agents in a sandbox town (Park et al., 2023) to million-agent replicas of social platforms (Yang et al., 2024), with intermediate steps prototyping communities (Park et al., 2022), benchmarking social intelligence (Zhou et al., 2024), and replicating human studies (Aher et al., 2023). Activation in these systems is handled by time-driven scheduling: OASIS's time engine assigns hourly activity probabilities with random draws, and agent activity levels are fixed at initialization. A complementary line attacks the scaling problem directly,

reaching millions of LLM-guided agents through archetype-based execution (Chopra et al., 2025) — again without changing *who* gets activated. Our work is orthogonal: we study the activation decision itself as a behavioral mechanism and show its design space is consequential and poorly charted. Validating generative social simulation against real data is a recognized difficulty (Tomašević et al., 2025); we contribute a validation layer for activation: responder-set matching against real cascades.

Information diffusion and cascade prediction. Predicting information cascades has a long history: size prediction from early adopters and content features (Cheng et al., 2014), and influence modeling under independent-cascade and linear-threshold models (Kempe et al., 2003). We differ in locus: rather than aggregate cascade outcomes, we predict the *participant set* given an event — the quantity a simulator must sample at run time.

Retrieval, negative sampling, and imbalance. Dense retrievers, from supervised bi-encoders (Karpukhin et al., 2020) to weakly supervised pretraining (Wang et al., 2022), provide strong off-the-shelf semantic ranking; collaborative filtering with implicit feedback relies on sampled negatives (Rendle et al., 2009); and negative-sampling ratio is known to shape representation quality in word embeddings (Mikolov et al., 2013) and beyond. Class imbalance is classically handled by resampling (Chawla et al., 2002) or loss reweighting (Lin et al., 2017); our inverted-U is a complementary, deployment-facing observation: the *sampling ratio* itself interacts with the gap between training and full-pool evaluation. Our setting recombines these threads in a simulation-facing regime — the candidate pool is the full agent population, the deployed positive rate is extreme ($\sim$1:1,750) — and we show that here negative-sampling decisions interact with full-pool evaluation in ways that held-out metrics systematically miss.

Evaluation methodology. Our paired design uses Wilson score intervals and exact McNemar tests (Wilson, 1927; McNemar, 1947). The distinction we stress — evaluation noise vs. training-run variance — is rarely stress-tested; our 36-run sweep quantifies both and uncovers failure modes invisible to held-out metrics.

3 Problem Formulation

Setting. An event stream delivers events $e \in \mathcal{E}$, each with text content c_e and arrival time t_e. A simulator maintains an agent pool $\mathcal{A} = \{a_1, \ldots, a_N\}$ where each agent a is instantiated from a real user's profile and historical posts H_a. When event e arrives, an *activation policy* selects a subset $S_e \subseteq \mathcal{A}$ of agents who are "woken up" and allowed to generate responses; all downstream simulation — reply content, cascade growth — unfolds over S_e.

Activation as ranking. We model activation as scoring every agent–event pair,

$$\pi_\theta(a \mid e) = P(\text{participate}_a = 1 \mid c_e, H_a, \text{profile}_a), \tag{1}$$

followed by a selection rule. Two selection forms matter in practice:

- **Match-size (top-k):** S_e = the top-$|C_e|$ agents by π_θ, where $|C_e|$ is the real cascade size. This is a *known-budget diagnostic* — it grants every policy the true responder count as budget, isolating ranking from size prediction; $|C_e|$ is unknown at deployment.
- **Threshold:** $S_e = \{a : \pi_\theta(a \mid e) > 0.5\}$. This form requires a probability with a *usable operating point*; it preserves endogenous cascade-size variability but exposes the policy's probability mass directly.

Ground truth and leakage control. For each test event e (user-based holdout, §5), the real responder set C_e is extracted from the observed cascade. Agent histories H_a are strictly truncated before t_e; the event author is never scored as a candidate responder; any reply text of the current event is scrubbed from histories. An early audit caught 90.8% of positive pairs leaking reply content into histories; the fix brought leakage to 0.0% and held-out F1 from an inflated 0.89 to its true 0.75 (Appendix B).

Table 1: The activation policy design space with measured costs. *Training*: one-time cost on our data. *Inference/event*: amortized cost to score the 6,566-agent pool for one event. *Operating point?* = whether the policy's scores support a usable fixed threshold for selection.

#	Family	Example	Training	Inference/event	Operating point?
F1	Random / popularity	uniform, activity-prior	none	negligible	trivially
F2	Time-engine scheduling	OASIS-style (Yang et al., 2024)	none	negligible	no (event-blind)
F3	Heuristic combination	recency+topic+activity	none	ms	no
F4	Dense retrieval	e5-base-v2 (Wang et al., 2022)	none (frozen)	ms (embed + dot)	**no (rank-only)**
F5	Collaborative filtering	matrix factorization (Rendle et al., 2009)	minutes	ms	no
F6	Discriminative cross-encoder	BERT pair classifier (Devlin et al., 2019)	$\sim$1.5 GPU-h	$\sim$4 s (full pool)	**yes**
F7	LLM judgment	zero-shot API scoring	none	API ($ / latency)	unclear
F8	Hybrid: retrieve + rerank	e5 shortlist + F6	inherits F6	$\sim$0.06 s	**yes** (inherited)

Deployment regime. The evaluation regime mirrors deployment: the *full* agent pool ($N = 6{,}566$) is scored for every test event, and the natural positive rate is extreme — an average event draws 3.7 unique responders from the pool, an effective ratio of $\sim$1:1,750 positives-to-candidates. This regime, not the balanced pair classification of training, is what an activation policy must survive.

Metrics. We report Precision@k (k=1, 3, 5) under match-size selection as the primary ranking metric; normalized F1 (nF1) for set quality; and for the threshold form, the *size bias* $|\bar{S}_e|/|\bar{C}_e|$ at a fixed 0.5 threshold (an operating-point probe, not distributional calibration — even calibrated probabilities need not concentrate above 0.5). All comparisons are paired per event: Wilson score intervals for single policies and exact McNemar tests for policy pairs (Wilson, 1927; McNemar, 1947).

4 The Activation Policy Design Space

We map activation policies into eight families spanning training cost, inference cost, and operating-point support. Table 1 summarizes; we instantiate and evaluate seven of them, deferring F7 for its API cost.

Three observations structure the space. First, activation is a *retrieval-flavored* problem, which is why F4 is a serious contender despite zero task training. Second, only F6 natively produces scores with a usable operating point for threshold-form simulation — a capability we show is decisive (§6.5). Third, F7 is the family most papers reach for first; our protocol admits an LLM row without redefining any metric (§8).

The discriminative family (main line). Our F6 instantiation is a BERT cross-encoder over the pair `[CLS] event [SEP] user-text [SEP]`, where user text concatenates profile fields, topic affinities, and recent posts. Training pairs are built from real cascades: each event–responder positive is paired with r negatives — the *negative-sampling ratio* r whose systematic study drives §6.1. Negatives are either random pool users or *hard* negatives: users who verifiably *saw* the event — they liked the post — but did not respond (§6.3).

5 Experimental Setup

Data. Our main testbed is a frozen Mastodon dataset: 5,411 seed posts, 5,260 reply cascades, 6,566 user profiles with posting histories. Cross-platform validation uses Zhihu, a Chinese Q&A platform with 450 questions and 2,840 users. Table 2 summarizes both platforms. Agent profiles are extracted from real behavior only — activity level, active hours, topic affinities — with no LLM-generated attributes. Beyond cascades, we collected the follow graph (7,233 users and 59,656 edges), boost chains, and *favorites* — the platform's like signal — which provide verifiable exposure events (§6.3).

Full-pool evaluation protocol. The split is user-based (80/20, shared across all conditions): a positive (event, responder) pair enters the test set iff its responder is held out, giving 2,241 test events; events may span both splits through different responders, so train-positive pairs are masked at evaluation. For each event we score the *entire* 6,566-agent pool. Every model dumps a full score matrix over all test events and agents; all metrics are computed offline from the matrix, with dump

Table 2: Platform characteristics. Zhihu cascade statistics are computed on the 88 fully-crawled threads.

	Mastodon	Zhihu
Language	English	Chinese
Events (seeds)	5,411	450
Reply cascades	5,260	88
Agent profiles	6,566	2,840
Size mean/med/p90	4.7/3/9	261/154/701
Max depth (mean)	2.3	2.5
Follow graph	7233/60k edges	—
Exposure signal	favorites	—

fidelity verified against the live policy (Reproducibility Statement). Train-positive pairs are masked at evaluation.

Multi-seed protocol. Motivated by the variance findings of §7, all headline numbers come from multiple independent runs varying both the negative-sample draw and training seed (s42/s43/s44). We report per-seed values and means; paired comparisons always use *same-seed* pairs.

Models. F6: `bert-base-uncased` cross-encoder over event–user-text pairs. F4: `intfloat/e5-base-v2` with `query:/passage:` prefixes, cosine similarity. F3: the best heuristic combination of semantic, recency, and activity signals from a 32-cell sweep. F2: an OASIS-style time-engine replay. F5: matrix factorization on the user–event participation matrix. Hyperparameters are listed in Appendix D. A full score matrix costs ∼2.6 GPU-hours with pre-tokenized caching.

Restricted-pool protocol. For cross-platform checks we also use the original restricted-pool protocol, in which the candidate set is the event's observed participant neighborhood; this isolates ranking quality from pool-size effects.

6 Results

The results below read the design space through the train/deployment distribution-gap lens: §6.1 varies negative-sampling *intensity*, §6.3 varies negative *source*, §6.4–6.5 separate ranking from operating-point behavior, and §6.6 assembles the remedy.

6.1 The inverted-U, and the catastrophic right end

Figure 1 shows full-pool match-size P@1 of the F6-family classifier as a function of the training negative-sampling ratio r, over 12 ratios from 1:1 to 1:500 with up to three seeds per point (per-seed values in Table 3).

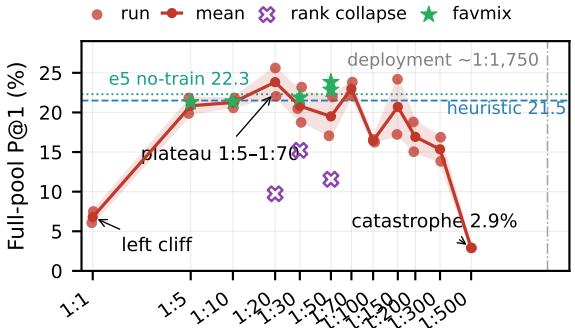

Figure 1: Full-pool P@1 vs. training negative-sampling ratio (log scale, 30 runs). 1:500, the tested point nearest the deployment ratio (∼1:1,750), yields 2.9%.

Three robust landmarks survive multi-seed scrutiny.

(i) The left cliff: balanced 1:1 training — the default in classifier tutorials — collapses to 6–7.5% under full-pool ranking. **(ii) A broad plateau** at 1:5–1:70 (20–23%): the optimal training ratio is one to two orders of magnitude *below* the deployment ratio, and the plateau is wide enough that exact tuning is unnecessary (its two † cells are ranking collapses, not ratio effects; §7). **(iii) The catastrophic right end**: at 1:300 the two surviving runs score 13.8% and

Table 3: Full-pool match-size P@1 (%) across training negative ratios and seeds (30 runs). †: ranking collapse (normal held-out F1, cliff-level P@1). ‡: all-negative attractor (F1 $= 0$; no checkpoint written, ranking unknown — §7). §: near-collapse (F1 $= 0.002$, ranking intact).

Ratio	seed 42	seed 43	seed 44	note
1:1	7.5	6.1	—	balanced-training cliff
1:5	21.8	19.9	—	plateau
1:10	21.4	21.8	20.6	plateau (most stable, ±0.6pp)
1:20	22.0	25.6	9.7†	plateau + collapse case
1:30	23.2	18.7	20.5	plateau
1:50	22.0	17.1	11.6†	plateau + collapse case
1:70	23.8	22.1	—	plateau right edge
1:100	16.2	16.6	—	degradation established
1:150	17.2	24.2	—	high-variance zone
1:200	18.8	15.0	—	high-variance zone
1:300	13.8	‡	16.9§	collapse region
1:500	**2.9**	‡	‡	**near-deployment catastrophe**

Table 4: All instantiated policy families under the full-pool protocol (2,241 events × 6,566 agents, match-size selection). P@1/P@3/P@5 and nF1 in %. F6 is the 1:50 configuration; F8 reranks the e5 top-100 with the F6 scorer.

#	Family	P@1	P@3	P@5	nF1	Size@0.5?
F1	Random	0.0	—	—	—	trivially
F1	Popularity	0.04	—	—	—	trivially
F2	Time-engine replay	0.00	0.01	0.05	0.14	no
F3	Heuristic (combined)	21.5	—	—	—	no
F4	e5 dense retrieval	22.3	11.6	7.5	23.2	**no (rank-only)**
F5	Matrix factorization	0.06	0.04	0.03	0.05	no
F6	BERT cross-encoder	22.0	11.7	9.3	—	**2.1**×
F8	e5 + BERT rerank	**22.4**	10.3	7.9	**25.1**	**2.1**× (inherited)
Oracle union (F4 ∪ F6)		29.4	—	—	—	—

16.9% while the third seed hits the all-negative attractor, and at 1:500 — still 3.5× below the true deployment ratio (∼1:1,750) — two of three seeds converge to the all-negative attractor (one on three consecutive attempts) and the sole survivor scores **2.9%**, below even the 1:1 cliff. "Match the training distribution to deployment" points past the worst configuration we measured.

6.2 The policy landscape under full-pool evaluation

Table 4 places all instantiated families under the same full-pool protocol; Figure 2 shows precision at k=1, 3, 5 for the competitive ones. Every policy is scored on the same 2,241 events against the same 6,566-agent pool; F7 (LLM-judgment) is part of the design space but its full-pool sweep is deferred (Limitations).

The landscape divides into two tiers. The bottom tier — policies that ignore event content (F1, F2) or cannot handle cold-start events (F5) — sits below 0.1%. The competitive tier (F3, F4, F6, F8) clusters at 21.5–22.4% P@1 — remarkably, spanning four orders of magnitude in training cost. What separates them is not ranking quality but everything else: a usable operating point, offered only by F6 and F8; deeper-k precision, where F6 wins P@5 at 9.3% vs. e5's 7.5%; and complementarity — the oracle union reaches 29.4%, so the competitive policies are far from redundant. No policy comes close to saturating the task.

6.3 Negative-sample quality beats ratio tuning — at the widest ratios tested

Favorites — Mastodon's like signal — give us something rare: *verifiable exposure without response*. A user who liked the seed post but never replied is a hard negative: exposed, interested, yet not a responder. Whether they help by exposure realism or sheer hardness is beyond our design to separate; here we test the empirical effect, replacing random negatives with these users — a mix we call favmix, holding all else fixed.

Table 5: favmix (verified-exposure hard negatives) vs. random negatives, same-seed paired (match-size P@1, %). All favmix runs are shown. †: ranking collapse (held-out F1 normal; Appendix C).

Ratio (seed)	favmix	random	Δ	McNemar p
1:5 (s42)	21.3	21.8	−0.5	0.36
1:10 (s42)	21.4	21.4	+0.0	1.00
1:30 (s42)	15.2†	23.2	−8.0	—
1:30 (s43)	21.9	18.7	+3.1	0.0002
1:50 (s42)	23.8	22.0	+1.9	0.0077
1:50 (s43)	22.9	17.1	+5.8	<0.0001

Two effects, both decision-relevant (Table 5). **Mean effect:** at 1:50, favmix beats its same-seed random baseline in both seeds (+1.9pp and +5.8pp). **Variance effect:** favmix is stable across seeds (Δ0.98pp, n.s.) while the random-negative twin swings by 4.9pp (p<0.0001) — hard negatives anchor the decision boundary where it matters. The advantage has a *quantity threshold*: nil at 1:5–1:10, fragile at 1:30 (one seed +3.1pp, one ranking collapse), reliable only at 1:50, the widest ratio tested.

6.4 Frozen dense retriever vs. trained classifier

The F4-family e5 retriever, with *no task training*, reaches 22.3% full-pool P@1 — no statistically detectable difference from the trained classifier's 22.0% (McNemar p=0.697; no significant difference detected). Two findings make this more interesting than a tie.

Complementarity is structural. The two models' wins are substantially complementary: their hit sets intersect at 14.9% against individual hit rates of ∼22% (per-event Spearman 0.148), and an oracle that gets credit when *either* is right reaches **29.4%**, +7pp over either alone. Slices are systematic (Table 7, Appendix D): e5 wins long-text and multi-participant events, where semantic signal is rich, while the trained classifier wins short-text events and, dramatically, singleton cascades (9.5% vs. 1.0%) — it has learned non-semantic regularities that embeddings do not carry.

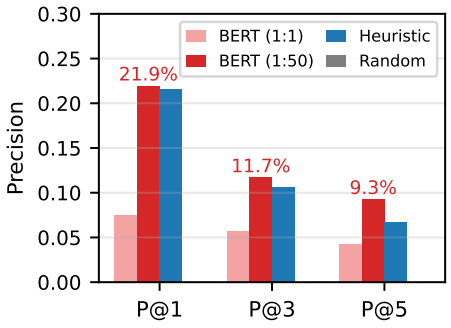

Figure 2: Ranking-head quality at k=1, 3, 5 (full pool, match-size). The balanced 1:1 model is broken at every k; the trained model separates at P@5.

Dead families are informative. The F2 time-engine replay scores 0.00% — scheduling without event semantics is vacuous under full-pool ranking — and F5 matrix factorization collapses to 0.06% under cold-start, since test events have no latent factors. These nulls bound the space from below.

6.5 Threshold behavior: rank-only vs. probability-bearing policies

The threshold form asks for probability *mass*, not just order. We probe this with the fixed-threshold size check of §3: agents selected at a 0.5 threshold, relative to the mean unique responder count (3.7 per test event). Along the inverted-U, the size bias falls monotonically with r (Figure 3): 320× over-selection at 1:1, crossing 1× between 1:50 (2.1×) and 1:100 (0.73×), to 0.39× at 1:200. Figure 4 makes the failure concrete: a 1:1-trained model's threshold form activates ∼1,198 agents per event against 3.7 real responders. Ranking quality and threshold-size behavior *decouple*: the near-unbiased ratios are not the best-ranked ones, and the 1:1 cliff is simultaneously a ranking failure and a 320× over-selection failure.

For e5, cosine similarities have no natural operating point: even after Platt scaling (2-fold cross-fitted — optimistic for the retriever), a 0.5 threshold selects 2,155 agents per event against 3.7 real responders (**575× size bias** vs. 2.1× for the 1:50 classifier). In our tests, dense retrieval is a rank-only policy: it can pick *who*, not *how many*; where a threshold form is needed, our evidence supports driving it with a discriminatively trained model.

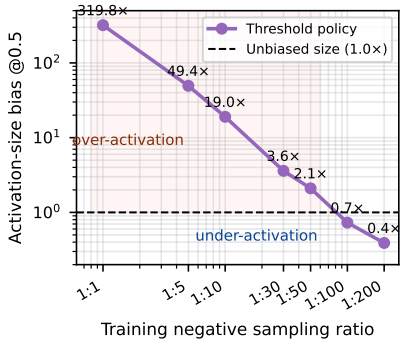

Figure 3: Activation-size bias at a fixed 0.5 threshold vs. training ratio (log–log). The bias crosses $1\times$ between 1:50 and 1:100.

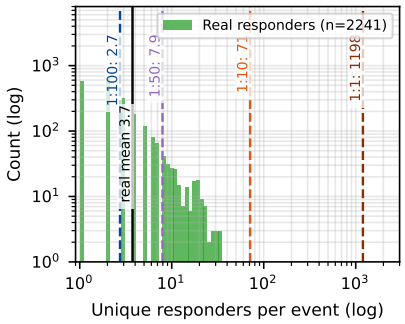

Figure 4: Real unique-responder distribution vs. fixed-threshold simulated means. Wrong ratios select one to two orders of magnitude too many agents.

6.6 A DEPLOYABLE HYBRID: RETRIEVAL RECALL + DISCRIMINATIVE RERANK

Full-pool cross-encoder scoring costs ∼2.6 GPU-hours per 2,241-event batch — affordable offline, expensive in an interactive loop. The complementarity of §6.4 suggests the standard two-stage answer: e5 retrieves a shortlist (top-100 recall: 35.0% of responders), the F6 classifier reranks it.

Reranking the e5 top-100 with the 1:50 classifier yields **22.4% P@1** — matching or exceeding both parents — while scoring only ∼100 of 6,566 candidates per event (∼1.5% of full-pool cost), and the hybrid's set-level quality is the best of any practical policy (nF1 25.1 vs. 23.2 for pure e5; Table 4). Because the final scores are discriminative, the hybrid inherits F6's threshold-size behavior — with the caveat that thresholding a shortlist caps recall at the retriever's ceiling. F8 is the configuration we recommend for interactive simulation: retrieval-speed inference with a usable operating point.

6.7 CROSS-PLATFORM CHECKS

The headline ordering replicates on a second platform: on Zhihu's restricted pool the trained classifier leads (0.956 P@1) with the frozen multilingual retriever close behind (0.900) and heuristics far below; on Mastodon's restricted pool the ordering is unchanged (BERT 0.626 > e5 0.590 > heuristic 0.589; Figure 5, Appendix I). Two platform-level asymmetries recur: Zhihu events are information-dense, with event text alone carrying most of the signal, while on Mastodon user profiles are necessary but bounded — removing them drops the classifier below the majority baseline (Appendix D).

Two cautionary findings shaped the Zhihu pipeline (Appendix I): its headline numbers sit on a structurally cold-started population (88.9% of users have no usable posting history), and an early negative-sampling artifact — positives largely history-less, random negatives all with history — produced an implausible F1 of 0.995: the model had learned "has a profile" rather than event semantics: negative-sampling bugs masquerade as breakthroughs.

7 EVALUATION VARIANCE: A METHODOLOGICAL NOTE

Metrics computed off the deployment distribution cannot see the failures that matter most, and *single-run* full-pool numbers are far less trustworthy than they look. Our seed-replication sweep (36 runs — 30 random-negative configurations plus 6 favmix — varying both negative draw and training seed) surfaces three phenomena that should change how activation policies are evaluated.

Run variance is large. Identical configurations retrained with a different seed swing by ±2–4.5 percentage points in full-pool P@1 (e.g. 1:50: 22.0% vs. 17.1% vs. 11.6%). Early single-seed conclusions in our own pipeline — a "peak" at 1:30, and an apparent right-tail "rebound" (1:200 > 1:100, p=0.002 paired) — both dissolved under replication (at a second seed the rebound reversed:

15.0% vs. 16.6%, p=0.092). A significant paired gap rules out *evaluation* noise only, not *training-run* variance: single-run comparisons within ∼5pp are inconclusive.

Held-out F1 and full-pool ranking decouple. At 1:30, the seed with the *better* held-out F1 (0.254 vs. 0.229) had the *worse* full-pool P@1 (18.7% vs. 23.2%): the two metrics stress different properties of the score head, and held-out F1 cannot proxy for simulation quality.

Two silent failure modes. Beyond smooth variance, we found two qualitatively distinct collapses, together affecting 6 of 36 runs — one in six (per-run ledger in Appendix C):

- **Ranking collapse** (3/36, ∼8%): held-out F1 stays normal (0.19–0.21) while full-pool P@1 crashes to cliff level (∼10%) — *invisible to training logs and held-out metrics*; the run looks healthy until deployed against the full pool.
- **All-negative attractor** (3/36, all at extreme ratios ≥1:300): training converges to the *predict-all-negative attractor* — post-mortem on a surviving 1:300 checkpoint shows majority-rate accuracy with near-zero recall, yet AUC 0.73 and a usable full-pool P@1 of 16.9%. The standard class-imbalance fix does *not* rescue it: reweighting positives by 50× or 500× still converges to the same attractor (F1 = 0) while the encoder keeps learning (AUC 0.70–0.81) — the binding constraint appears to be the *temporal sparsity* of positives (each seen once, in ∼6% of batches) under one-epoch training, not gradient imbalance. The attractor is near-deterministic at 1:500 (seed 44: 3/3 attempts; seed 43: 1/1). And because no checkpoint is written while the log still prints a plausible "best model" path, pipelines keying success on exit codes silently record these runs as successes.

Protocol recommendation. Our resulting protocol, used for every number in §6: replicate every point with ≥2 seeds, report per-seed values alongside means, pair comparisons within the same seed, treat McNemar significance as a test of evaluation noise (not replicability), and run a full-pool sanity check before trusting any run. The cost is modest next to the cost of publishing a seed artifact.

8 Discussion and Conclusion

Practical guidance. Our results suggest a decision procedure for builders of agent-based social simulators. (i) *If you cannot train*: use a dense retriever (F4): ranking matches a trained classifier at zero cost — but only for match-size selection; never threshold raw similarities. (ii) *If you can train* a discriminative cross-encoder (F6): train on the 1:5–1:70 plateau, one to two orders below the deployment ratio: ranking is best and the 0.5-threshold size bias is within a factor of two. (iii) *Invest in negative quality, not ratio tuning*: verified-exposure hard negatives beat their same-seed random twins at 1:50 in both seeds and tame run-to-run variance. (iv) *Evaluate the way you deploy*: full-pool ranking on multiple seeds; held-out F1 is a training diagnostic, not evidence of simulation quality.

Limitations. Our main evidence comes from a single platform, Mastodon; Zhihu checks are restricted-pool and smaller. The F7 LLM-judgment sweep is deferred (per-event API cost at full-pool scale). The classifier is a modest BERT-base; whether the inverted-U persists at larger scale is open, though the driver — the train/deployment distribution gap — is architecture-independent. Three caveats bound our claims: the size-bias probe is an operating-point check, not distributional calibration; favmix's gain cannot be attributed to exposure realism versus hardness within our design; and deployment-ratio training is an extrapolation, not a measurement — our rightmost point is 1:500, well short of ∼1:1,750.

Conclusion. We recast agent activation in LLM social simulation as a behavioral mechanism with its own design space, and mapped that space under a deployment-faithful, full-pool protocol. The landscape is consequential and treacherous: a broad sweet spot far below the deployment ratio, collapse far short of it, negative quality dominating quantity, retrieval that ranks but cannot set its own operating point, and failure modes invisible to held-out metrics. Activation deserves first-class treatment — and what it is trained and evaluated on matters as much as the model: close the gap where it is cheap, and evaluate where you deploy.

AI USE STATEMENT

In this work, we used generative AI tools for implementing methods and experiment code, exploring experimental designs, creating and editing scientific figures, and drafting and editing this paper, including assistance in organizing the presentation of results. We did not use generative AI tools to generate datasets. All scientific claims, experimental designs, and results were decided, verified, and are endorsed by the authors, who take full responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

All data are public posts from Mastodon and Zhihu; agent profiles contain only public behavioral aggregates, and we release no re-identifiable content. Simulation is not a substitute for human-subject studies.

REPRODUCIBILITY STATEMENT

All evaluation numbers are computed offline from dumped score matrices, so every metric, pairing, and significance test in the paper can be recomputed exactly; the frozen dataset snapshot, seeds, hardware, and hyperparameters are specified in Appendix A, with the leakage-control audit in Appendix B. Training code, matrix-evaluation code, queue manifests with per-run seeds, and all score matrices will be released upon publication.

REFERENCES

Gati V. Aher, Rosa I. Arriaga, and Adam Tauman Kalai. Using large language models to simulate multiple humans and replicate human subject studies. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, pp. 337–371, 2023.

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

Justin Cheng, Lada Adamic, P. Alex Dow, Jon Michael Kleinberg, and Jure Leskovec. Can cascades be predicted? In *Proceedings of the 23rd International Conference on World Wide Web (WWW)*, pp. 925–936, 2014.

Ayush Chopra, Shashank Kumar, Nurullah Giray Kuru, Ramesh Raskar, and Arnau Quera-Bofarull. On the limits of agency in agent-based models. In *Proceedings of the 2025 International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2025.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 4171–4186, 2019.

Chen Gao, Xiaochong Li, Nian Jin, Yisong Song, et al. S^3: Social-network simulation system with large language model-empowered agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP) System Demonstrations*, 2023.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6769–6781, 2020.

David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 137–146, 2003.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2980–2988, 2017.

Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26 (NeurIPS)*, pp. 3111–3119, 2013.

Joon Sung Park, Lindsay Popowski, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Social simulacra: Creating populated prototypes for social computing systems. *Proceedings of the ACM on Human-Computer Interaction (CSCW)*, 6, 2022.

Joon Sung Park, Joseph O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST)*, 2023.

Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. BPR: Bayesian personalized ranking from implicit feedback. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 452–461, 2009.

Aleksandar Tomašević, Darja Cvetković, Sara Major, Slobodan Maletić, Miroslav Anđelković, Ana Vranić, Boris Stupovski, Dušan Vudragović, Aleksandar Bogojević, and Marija Mitrović Dankulov. Towards operational validation of LLM-agent social simulations: A replicated study of a Reddit-like technology forum. *arXiv preprint arXiv:2508.21740*, 2025.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*, 2022.

Edwin B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927.

Ziyi Yang, Zaibin Zhang, Zirui Zheng, Yuxian Jiang, Ziyue Gan, Zhiyu Wang, et al. OASIS: Open agent social interaction simulations with one million agents. *arXiv preprint arXiv:2411.11581*, 2024.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. SOTOPIA: Interactive evaluation for social intelligence in language agents. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

A Reproducibility Statement

All evaluation numbers are computed offline from dumped score matrices (2241×6566 fp32 per run), so every metric, pairing, and significance test in the paper can be recomputed exactly from the matrices without re-running models. Matrix dumps pass an equality check against the live policy (30 random pairs including cross-batch-boundary cases, max diff $< 2 \times 10^{-3}$). All Mastodon experiments use the same frozen dataset snapshot (internal tag `v20260515`) and a fixed split seed (42) shared across all conditions; training runs vary negative-draw seed and training seed (s42/s43/s44). All training and scoring runs use RTX 4090 GPUs. Training code, matrix-evaluation code, queue manifests with per-run seeds, and all matrices will be released upon publication.

B Leakage Control Audit

An early version of the pipeline leaked reply content: agent histories H_a included the agent's own reply to the *current* event, so the classifier could match reply text to event text (90.8% of positive pairs affected; held-out F1 inflated to 0.89). The triple fix: (i) histories strictly truncated at t_e; (ii) event author excluded from candidates; (iii) current-event reply text scrubbed from all histories. Post-fix audit: 0.0% measured leakage; held-out F1 0.75; all numbers in this paper are post-fix.

Table 6: The six collapsed runs out of 36 training runs (17%; seed 44 at 1:500 additionally collapsed on two further attempts, 3/3 in total). Ranking-collapse runs show a *normal* held-out F1 while full-pool P@1 falls to cliff level — invisible to training logs. All-negative-attractor runs converge to predicting no positives (F1 $= 0$) while the encoder keeps learning (AUC 0.70–0.81); because training saved checkpoints only on held-out F1 improvement, these runs wrote no model, so their full-pool ranking quality is unknown. The training script now always saves the final model.

Config	Held-out F1 / AUC	Full-pool P@1	Failure type
1:20 random, s44	0.19–0.21 / 0.83–0.85	9.7%	ranking collapse
1:50 random, s44	0.19–0.21 / 0.83–0.85	11.6%	ranking collapse
1:30 favmix, s42	0.19–0.21 / 0.83–0.85	15.2%	ranking collapse
1:300 random, s43	0.0000 / 0.70–0.81	— (no checkpoint)	all-negative attractor
1:500 random, s43	0.0000 / 0.70–0.81	— (no checkpoint)	all-negative attractor
1:500 random, s44	0.0000 / 0.70–0.81	— (no checkpoint)	all-negative attractor

C Failure-Mode Ledger

Post-mortem on the one attractor-adjacent run that did write a checkpoint (1:300, s44, near-collapse with F1 $= 0.002$): majority-rate accuracy 0.9961 with recall 9×10^{-4}, yet AUC 0.73 and a usable full-pool P@1 of 16.9% — the encoder ranks correctly while the fixed-threshold classifier has collapsed onto all-negative.

D Additional Experimental Details

Training hyperparameters. `bert-base-uncased` cross-encoder; input `[CLS] event-text [SEP] user-text [SEP]`; max length 256; 1 epoch; AdamW lr 2×10^{-5}; batch 32; dropout 0.3; weight decay 0.01; best checkpoint by held-out F1. A full 12-ratio curve plus score-matrix dump costs $\sim$2.6 GPU-hours per point on an RTX 4090 with pre-tokenized caching.

Statistical tests. Single-policy intervals: Wilson score interval on per-event Bernoulli precision (Wilson, 1927). Paired comparisons: exact two-sided McNemar test on per-event discordant pairs (McNemar, 1947). Spearman rank correlation for per-event score agreement between families.

Table 7: Complementarity slices (full-pool match-size P@1, %). Oracle credits an event if either policy's top-1 is a true responder.

Slice (n)	F6	F4	Oracle
size $\leq$1 (589)	**9.5**	1.0	10.2
size 2 (650)	30.9	**38.9**	44.5
size 3 (316)	23.4	**28.2**	33.2
size 4–5 (303)	**27.1**	25.7	33.3
size $>$5 (383)	**20.6**	19.3	26.9
short text (1121)	**21.3**	16.7	25.1
long text (1120)	22.6	**27.9**	33.6

Experiment index. Table 8 maps paper claims to internal experiment IDs for traceability.

Table 8: Claim-to-experiment index.

Claim	Experiment IDs
Inverted-U curve (12 ratios)	EXP-2026W30-016, EXP-2026W31-001
Seed replication / variance	EXP-2026W30-017, 018
favmix paired effects	EXP-2026W30-019
e5 head-to-head & oracle	head2head matrix suite (2026-07)
Threshold size bias	EXP-2026W30-016 (threshold form)
Cross-platform (Zhihu, restricted pool)	restricted-pool suite (2026-06)

E Profile Component Ablation

Held-out pair classification on the frozen Mastodon dataset (leakage-fixed protocol), varying which profile components the classifier sees. Majority baseline F1 $= 0.6729$, random F1 $= 0.5161$.

Table 9: Profile component ablation (held-out F1/AUC, Mastodon, frozen dataset).

Variant	F1	AUC
Full profile	0.702	0.824
Content only (no structured fields)	0.718	0.819
Tags only	0.707	0.806
No history (empty profile)	0.646	0.466
Majority baseline	0.673	0.500

The decisive comparison is full vs. empty profile: removing the user history drops F1 *below the majority baseline* (0.646 vs. 0.673, AUC 0.466 — worse than random), confirming that the model's signal genuinely comes from the profile rather than dataset artifacts. The profile's contribution is platform-dependent: $+0.056$ F1 on Mastodon (necessary but bounded) vs. $+0.254$ on Zhihu (information-dense events still leave much to profiles).

F Heuristic Baseline Sweep

Seven heuristic scorers under held-out pair classification (Mastodon and Zhihu), min-max normalized, with best-threshold F1* and F1 at 0.5.

Table 10: Heuristic baselines (held-out). Zhihu heuristics degenerate (missing activity/history fields), which itself motivated the learned models.

Mastodon	AUC	F1@0.5	F1*
tfidf similarity	0.719	0.146	0.678
combined (6 weighted)	0.703	0.180	0.700
history match	0.631	0.205	0.678
time match	0.579	0.013	0.678
topic overlap	0.550	0.010	0.678
popularity	0.484	0.020	0.678
activity	0.475	0.001	0.678
Zhihu	**AUC**	**F1@0.5**	**F1***
all seven heuristics	0.44–0.50	0.00–0.09	0.667 (majority)

On Mastodon, single-feature heuristics reach AUC 0.55–0.72; the combined heuristic's full-pool P@1 is 21.5% (main text, Table 4). On Zhihu every heuristic collapses to the majority baseline — the platform lacks activity-level and active-hours fields; a Chinese-configured tfidf similarity recovers useful signal (AUC 0.860) but remains far below the learned classifier — which is precisely the gap learned models and multilingual embeddings fill.

G Who Responds: Graph Structure of Responders

Do responders differ structurally from non-responders in the follow graph (7,233 users, 59,656 edges; 4,676 responders vs. 2,497 non-responders)?

Table 11: Graph centrality of responders vs. non-responders (median values; KS test against the two distributions).

Metric	Responders	Non-responders	Effect ratio	KS p
In-degree (followers)	2	1	2.0	$< 10^{-5}$
Out-degree (following)	1	1	1.0	$< 10^{-7}$
Degree	4	3	1.3	$< 10^{-5}$

Responders are significantly better connected (median in-degree 2 vs. 1), but the effect sizes are small — graph position alone is a weak activation signal, consistent with popularity scoring 0.04% P@1 in the full-pool protocol (Table 4).

H RESPONDER-SET DISTRIBUTION MATCHING

A restricted-pool sanity task (74 events): predict the responder *set* and compare to the real cascade, measuring set F1 and Jaccard.

Table 12: Responder-set matching (74 events, restricted pool).

Policy	Set F1	Jaccard	Precision
BERT (trained)	0.742	0.697	0.751
Majority	0.667	0.500	0.500
Random	0.504	0.442	0.504

I ZHIHU STRUCTURAL COLD START

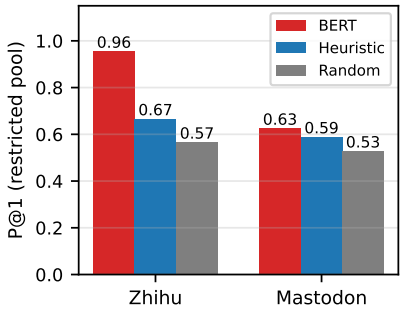

Figure 5: Restricted-pool match-size P@1 on both platforms. Family ordering is stable; absolute numbers exceed full-pool values because the candidate set is restricted.

Of Zhihu's 2,840 users, 2,526 (88.9%) have no usable posting history (fallback profiles). Split by profile provenance: 99.7% of the 313 high-quality (profile-page) users have histories, vs. 0.1% of the 2,527 fallback users. A naive profile-quality comparison is therefore confounded by sample size ($\sim$4,300 vs. $\sim$514 training pairs), and an early configuration produced F1 0.995 through negative-sampling bias (positives mostly history-less, negatives all with history) — fixed by stratifying the negative draw on history availability. These structural facts are why we treat Zhihu as a restricted-pool robustness check rather than a second full-pool benchmark.