Tiny Society AI: Calculations and Measurements

A practical guide to the model's formulas and update rules

Tiny Society AI
Multi-Agent Social Simulation and Prediction
16 July 2026

Abstract

This paper explains how Tiny Society AI turns agent actions into the numbers shown during a simulation and in the final forecast. It covers relationships, shared beliefs, memory, feeds, agent selection, influence, and stance changes. Each section first explains what a value means, then gives its formula and valid range. The aim is to make the model easy to inspect and reproduce. Unless noted otherwise, one simulated "day" includes morning, afternoon, evening, and night.

1 Scope and notation

Let $\mathcal{A}$ denote the set of agents, $N = |\mathcal{A}|$ the population size, $t \in \{0, 1, \ldots, T\}$ the simulated day, and $\mathcal{T}$ the set of stance topics. A directed relationship from agent i to agent j has a categorical type and signed strength r_{ij}. The main number ranges are:

Quantity	Symbol	Domain
Relationship or stance strength	r_{ij} or $s_{i,q}$	$[-1, 1]$
Influence score	I_i	$[-100, 100]$
Memory importance	m_i	$[1, 10]$
Simulated day	t	one morning–night tick

Relationships have a direction. Agent i can have a relationship with agent j even when the reverse relationship does not exist, and the two directions can have different labels. A category count includes the pair once if the label appears in either direction. The operator $\mathrm{clamp}(x, a, b) = \min(\max(x, a), b)$ keeps x between a and b.

Measurement convention. "Meaning" explains a measure in plain language. "Definition" shows how the code calculates it. A listed range is the mathematical range unless the text gives a smaller cap.

1.1 Connection to Stanford's Generative Agents paper

Tiny Society AI takes inspiration from Stanford's *Generative Agents* study by Park et al. [1]. Both systems let agents store experiences as text, retrieve memories using relevance, recency, and importance, form reflections, and use remembered information when planning and interacting. This project is not a reproduction of that system. It adds directed relationships, topic stances, group-level measures, forecast scores, and fixed update rules. It also measures memory relevance with token overlap instead of embeddings, uses its own recency factor, and avoids returning several near-duplicate memories. The Stanford paper provides the starting idea; the formulas below describe this project's implementation.

2 Social metrics

For each unordered agent pair $p = \{i, j\}$, let T_p be the set of categories appearing in either stored direction. The count for relationship category c is

$$C_c = \sum_{p \in \mathcal{P}} \mathbb{1}[c \in T_p], \tag{1}$$

where $\mathcal{P}$ contains every pair with at least one relationship. A one-way relationship counts once, and a two-way relationship also counts once.

Table 1: Daily macro metrics.

Measurement	Meaning	Definition	Range
friendship_count, rivalry_count, conflict_count, romance_count, alliance_count	Number of extant relationships of each type	Equation 1 for the corresponding type	≥ 0
average_relationship_strength	Average bond strength, ignoring whether it is positive or negative	$\frac{1}{\lvert E\rvert}\sum_{(i,j)\in E}\lvert r_{ij}\rvert$	$[0,1]$
average_trust_score	Average positive strength of trust-labelled edges	Mean of r_{ij} over $E^{+}_{\text{trust}} = \{(i,j) : c_{ij} = \text{trust},\ r_{ij} > 0\}$; returns 0 when the set is empty	$[0,1]$
most_connected	Agents with the most outgoing relationships	Top three agents by outgoing relationship count	names
relationship_volatility	Significant realized relationship transitions during the day	Number of milestone events emitted by the consequence layer; see Section 7	≥ 0
social_fragmentation	Fraction of isolated agents	$\lvert\{i : \deg^{+}(i) = 0\}\rvert / N$	$[0,1]$
group_centrality	Groups with the most links to outsiders	Per-group count of directed edges from members to non-members; five highest retained	counts
influence_gainers, influence_losers	Agents whose standing rose or fell since initialization	Three highest/lowest values of $I_i(t) - I_i(0)$, subject to the appropriate sign	names

The system saves each agent's starting influence $I_i(0)$ once. Influence gainers and losers are therefore measured against Day 0, not against the previous day.

3 Group beliefs and forecast confidence

Each agent i holds a stance $s_{i,q} \in [-1,1]$ on topic $q \in \mathcal{T}$. For the subset $\mathcal{A}_q$ of agents holding topic q, the population mean is

$$\mu_q(t) = \frac{1}{\lvert\mathcal{A}_q\rvert}\sum_{i\in\mathcal{A}_q} s_{i,q}(t), \qquad \mu_q(t) \in [-1,1]. \tag{2}$$

The standard deviation measures how much agents disagree:

$$\sigma_q(t) = \sqrt{\frac{1}{\lvert\mathcal{A}_q\rvert}\sum_{i\in\mathcal{A}_q}\left(s_{i,q}(t) - \mu_q(t)\right)^2}, \tag{3}$$

If only one agent has a stance on the topic, $\sigma_q(t) = 0$. The code uses population standard deviation and rounds the mean and standard deviation to four decimal places.

The aggregate swarm confidence is

$$\Gamma(t) = \operatorname{clamp}\left(1 - \frac{1}{\lvert\mathcal{T}\rvert}\sum_{q\in\mathcal{T}}\sigma_q(t), 0, 1\right). \tag{4}$$

Interpretation. $\Gamma = 1$ means complete agreement. Lower values mean the agents are more divided. This is a measure of consensus, not the probability that the forecast is correct.

3.1 Final forecast object

The final forecast records the following quantities:

- `topic_means`, `topic_uncertainty`, and `confidence` are copied from the final day.
- A day's topic drift is

$$D(t) = \sum_{q\in\mathcal{T}}\left\lvert\mu_q(t) - \mu_q(t-1)\right\rvert. \tag{5}$$

Positive-drift days are ranked by $D(t)$; the three largest are retained as pivotal days (`MAX_PIVOTAL_DAYS`=3).

- `narrative` is produced by the single `FINAL_REPORT` language-model call from the belief trajectory and pivotal-day causal attributions.
- `question` carries the player's prediction question when one was supplied; it may be null in an open-ended sandbox run.

4 Memory retrieval

Following Park et al. [1], memory retrieval balances relevance, recency, and importance. The formulas below show this project's version of that idea.

4.1 Importance at creation

For memory text x, let $\ell(x)$ be its character length and $h(x)$ the number of emotional-keyword hits. Raw importance is

$$m_{\text{raw}}(x) = 1 + \min\left(\frac{\ell(x)}{60}, 4\right) + \min(1.5h(x), 5)\,, \tag{6}$$

$$m(x) = \text{clamp}(m_{\text{raw}}(x), 1, 10). \tag{7}$$

Reflections override the computed value with 9.0; whisper-advice memories use 10.0.

4.2 Retrieval score

Given a query, a current day, and default return count $k = 8$, each candidate memory j receives three component scores:

$$\rho_j = \frac{|Q \cap M_j|}{|Q|}, \qquad \text{stopword-filtered token overlap,} \tag{8}$$

$$\tau_j = 0.85^{t-a_j}, \qquad \text{recency since last access day } a_j, \tag{9}$$

$$\iota_j = \frac{m_j}{10}, \qquad \text{stored importance.} \tag{10}$$

Here Q is the query-token set and M_j the memory-token set. Each component is min–max normalized across the current candidate pool before aggregation. With all weights equal to one,

$$S_j = \widetilde{\rho}_j + \widetilde{\tau}_j + \widetilde{\iota}_j. \tag{11}$$

The system also tries not to return several memories that say the same thing. If C_m is the set already chosen at step m and $J(M_j, M_h) = |M_j \cap M_h| / |M_j \cup M_h|$ measures shared tokens, the next memory is

$$j^* = \arg\max_{j \notin C_m} \left[S_j - 0.6 \max_{h \in C_m} J(M_j, M_h) \right], \tag{12}$$

with no penalty when C_m is empty. The process stops after selecting k memories. When a memory is returned, its `last_accessed_day` is updated, making it recent again.

5 Feed routing and ranking

Each agent sees a different subset of the global activity stream.

5.1 Witness routing

Table 2: Audience rule by action type.

Action	Witness set
`post`	Entire population; a public broadcast.
`direct`	Actor and explicitly named targets only.
`comment`, `interact`, `amplify`	Actor, targets, and the actor's group-mates; additionally the entire population when the actor's influence is at least `PUBLIC_INFLUENCE_THRESHOLD`=20.0.

An `amplify` action also inserts a synthetic entry attributed to the amplified target into the audience's feed, propagating that target's standing. Each feed retains at most `OBSERVATION_WINDOW`=15 entries.

5.2 Ranked feed

For viewer v and entry e, let V_v be a token set assembled from the viewer's traits, revealed traits, stance topics, relationship names, and own name; let E_e be the entry tokens. The components are

$$\text{interest}(v, e) = \frac{|V_v \cap E_e|}{|V_v|}, \tag{13}$$

$$\text{influence}(e) = \frac{I_{\text{author}(e)}}{\max_{u \in \text{feed}(v)} \left| I_{\text{author}(u)} \right|}, \tag{14}$$

$$\text{recency}(e) = 0.85^{t_{\max} - t_e}. \tag{15}$$

The rank score is

$$F(v, e) = 3.0\ \text{interest}(v, e) + 1.0\ \text{influence}(e) + 1.5\ \text{recency}(e). \tag{16}$$

Shared interests have the largest effect on ranking. Author influence and newer posts have smaller effects.

6 Choosing active agents

Each agent i receives a selection score from four terms:

$$\begin{aligned} z_i = {} & 3.0\,\mathbb{1}[\text{event-proximate}_i] + 2.0 \frac{I_i}{\max_j I_j} \\ & + 1.5\,\mathbb{1}[\text{charged relationship}_i] + 0.5 \min(g_i, 3), \end{aligned} \tag{17}$$

where event proximity means that a token from a trait, goal, group, or name occurs in the day's event; a charged relationship is a rivalry, romance, or conflict with absolute strength at least 0.3; and g_i is the number of groups held by agent i.

Agents are sampled without replacement. At draw m, with remaining pool R_m, the conditional probability is

$$\Pr(i \text{ chosen at } m \mid R_m) = \frac{\exp(z_i/\theta)}{\sum_{j \in R_m} \exp(z_j/\theta)}, \qquad \theta = \texttt{SOFTMAX\_TEMPERATURE} = 1.0. \tag{18}$$

This is not a fixed top-N list, so the active agents can change from day to day. A lower temperature favors high scores more strongly; a higher temperature gives other agents a better chance. The default daily limit is eight agents, configurable from 1 to 30. When the limit is at least the population size, every agent acts.

7 Relationship updates

An agent's social intent changes a numeric affinity score. The code derives relationship labels from that score and the interaction history; the language model does not set the final label.

7.1 Intent values and accumulation

Each intent maps to a proposed relationship dimension and calibrated bid b:

Intent	**Dimension**	b	**Intent**	**Dimension**	b
befriend	friendship	+0.18	support	friendship	+0.15
comfort	friendship	+0.16	praise	friendship	+0.12
reconcile	friendship	+0.20	confide	trust	+0.18
trust	trust	+0.15	flirt	romance	+0.15
court	romance	+0.22	ally	alliance	+0.20
collaborate	alliance	+0.16	confront	conflict	−0.22
rebuke	conflict	−0.18	reject	conflict	−0.20
undermine	rivalry	−0.20	challenge	rivalry	−0.15
compete	rivalry	−0.12	distance	trust	−0.12
talk	trust	+0.05	observe	trust	+0.02
unknown/blank	trust	+0.05			

For a directional edge with affinity r_{ij}, the accumulated update is

$$r_{ij}(t+1) = \text{clamp}\big(r_{ij}(t) + 0.5\ \text{clamp}(b, -0.4, 0.4), -1, 1\big). \tag{19}$$

Each bid increments the directional interaction count. A romantic or alliance bid also adds 0.25 to the corresponding directional signal accumulator, capped at 1.0. A missing edge is initialized as trust with zero affinity before the update. The actor's bid is applied directly; the target's bid first passes through the perception layer in Section 9 and is then damped by Equation 19.

7.2 Derived relationship categories

After both directional accumulations, the model assigns the first qualifying category in Table 3. Here n_{ij} is the directional interaction count, u_{ij} the romantic signal, and a_{ij} the alliance signal. Near-neutral affinity that meets no rule retains its existing label to avoid category churn.

Table 3: Ordered relationship realization rules.

Assigned type	First qualifying condition
romance	$r_{ij}, r_{ji} \geq 0.50$; $u_{ij}, u_{ji} \geq 0.40$; and $n_{ij} \geq 4$
alliance	$r_{ij}, r_{ji} \geq 0.35$; $a_{ij}, a_{ji} \geq 0.40$; and $n_{ij} \geq 3$
conflict	$r_{ij} \leq -0.50$
rivalry	$r_{ij} \leq -0.25$
friendship	$r_{ij} \geq 0.30$
trust	$r_{ij} \geq 0.10$
unchanged	None of the preceding conditions

Romance and alliance require evidence in both directions and several interactions. Conflict and rivalry can be one-sided. Existing relationships start with enough matching affinity and history to keep their initial label.

7.3 Milestones and reported volatility

`relationship_volatility` counts meaningful changes, not every internal label update. A change counts when a new bond reaches $|r| \geq 0.2$, when a relationship moves between positive and negative families, or when an existing type crosses the strong-bond threshold $|r| = 0.6$. Switching between rivalry and conflict, or among friendship, alliance, and trust, does not count as a turning point.

8 Influence updates

Influence deltas are derived from action type rather than authored by an agent.

The actor receives one base increment per action: 0.6 for `post`, 0.3 for `amplify`, 0.2 for `comment` or `interact`, and 0.1 for `direct`; an unknown kind defaults to 0.2. If at least one target receives a negative affinity bid, the actor receives an additional 0.3 once, and each antagonized target receives -0.4. Thus

$$\Delta I_{\text{actor}} = k(\text{action}) + 0.3\,\mathbb{1}[\exists \text{ antagonized target}], \qquad \Delta I_{\text{target}} = -0.4\,\mathbb{1}[b < 0]. \tag{20}$$

An `amplify` action separately grants each valid amplified target +2.0. Every mutation is clamped to $[-100, 100]$. Background-agent influence later regresses by a factor of 0.99 per day (Section 11). Influence enters reasoning-agent selection (Section 6), feed ranking (Section 5), and the public-figure audience threshold.

9 How agents perceive actions

When one character acts on another, the recipient's personality and history can change the effect on their relationship. This extra perception step runs only when both conditions hold:

$$|\delta_{\text{raw}}| \geq 0.05 \quad \wedge \quad (\text{prior relationship } \vee \text{ memory mentions actor}). \tag{21}$$

Otherwise, the raw delta is applied unchanged. When invoked, a language-model call returns a perceived delta

$$\delta_{\text{applied}} = \text{clamp}(\delta_{\text{perceived}}, -1, 1), \tag{22}$$

together with a short explanation and, sometimes, a newly revealed trait. An agent can have at most ten revealed traits. This allows the same action to affect two recipients differently.

10 Stance updates

An action may contain `stance_shift` values $\Delta s_{i,q}$ for the acting agent. Each represented topic is updated by

$$s_{i,q}(t+1) = \text{clamp}\big(s_{i,q}(t) + \text{clamp}(\Delta s_{i,q}, -0.3, 0.3), -1, 1\big)\,. \tag{23}$$

The prompt restricts updates to topics engaged by the action; the applicator independently enforces the $[-0.3, 0.3]$ per-action shift bound.

11 Background-agent updates

Agents that are not selected for language-model reasoning still change through these fixed daily rules.

Table 4: Deterministic daily transitions.

Rule	Eligibility	Transition
Bond upkeep/fade	Every stored bond of a background agent	Positive affinity gains 0.02 when the agent is social, charismatic, loyal, or patient; otherwise positive affinity loses 0.01 and negative affinity gains 0.01 toward zero. Affinity is capped to $[-1, 1]$ and the category is re-derived.
Isolation penalty	Agent has an introverted, anxious, or lonely trait and no relationships	Influence -0.1 per day, floored at -10.
Shared-group trust	Background-agent/peer pass shares at least one group	With probability 0.08 on the eligible pass, both directions gain +0.01 affinity and one interaction; categories are re-derived.
Group stance drift	Agent has peers sharing a group and topic q	With peer mean $\bar{s}_{G,q}$, update $s_{i,q} \leftarrow \text{clamp}(s_{i,q} + 0.02(\bar{s}_{G,q} - s_{i,q}), -1, 1)$; reverse the delta for contrarian, rebellious, stubborn, skeptical, independent, or defiant traits.
Influence regression	Background agent has nonzero influence	$I_i \leftarrow 0.99 I_i$, applied after the isolation adjustment and rounded to two decimals.

12 Main parameters

Table 5 lists the main timing rules, thresholds, and limits.

Table 5: Principal cadence and cap constants.

Parameter	Value	Operational role
Daily reasoning capacity	8 (range 1–30)	Maximum number of language-model reasoning agents per day.
Reflection interval	4 days	Selected agents synthesize higher-order reflections every fourth day.
Plan refresh horizon	3 days	A short-term plan is renewed when it reaches this age.
Exchange-turn cap	2	Maximum additional turns in a charged interaction.
Daily exchange-chain cap	3	Maximum charged interaction chains during one simulated day.
Multi-turn gate	charged type, $\lvert r \rvert \geq 0.4$, or charged intent	Activates for rivalry, romance, conflict, a strong relationship, or an intent mapped to a charged category.
Vignette capacity and gate	2; probability 0.70	Limits theatrical moments and controls whether vignette generation is attempted.
Dynamic-event interval	every 5 days after Day 3	Introduces automatically generated mid-run events.
Initial-event window	Days 1–3	Duration for which the starting event remains active.
Observation window	15 entries	Feed entries retained per agent.
Public-influence threshold	20.0	Standing at or above which eligible actions become public.
Selection temperature	1.0	Controls selection randomness; larger values increase variety.
Recency multiplier	0.85	Multiplicative per-day decay for memory and feed recency.
Recall diversity penalty	0.6	Maximum Jaccard redundancy penalty in diversified recall.
Affinity damping / signal increment	0.5 / 0.25	Dampens each affinity bid and accumulates romance or alliance evidence.
Bond decay / warmth upkeep	0.01 / 0.02	Controls background movement toward neutrality and positive upkeep by social agents.
Stance conformity / influence retention	0.02 / 0.99	Sets group-mean stance drift and daily background influence retention.
Reflection / advice importance	9.0 / 10.0	Makes synthesized insights and player guidance highly salient in recall.
Generation batch size	3 agents	Number of filler agents generated per language-model call.
Pivotal-day count	3	Number of high-drift days surfaced in the forecast.
Inference concurrency	6 calls	Maximum simultaneous calls in the daily reasoning fan-out.

13 Order of operations

The formulas stay the same, but their order determines which version of the state they use.

1. **Concurrent inference.** Selected agents' planning and reasoning calls fan out with a concurrency bound of six simultaneous calls.
2. **Sequential mutation.** Returned actions are applied in selection order so relationship, influence, and stance mutations remain deterministic for a fixed set of generated outputs.
3. **Serial exchanges.** Multi-turn interactions remain serial because every response depends on the previously applied turn.
4. **Tiered model routing.** Routine reasoning, vignettes, filler generation, planning, and dynamic events use a cost-efficient model tier. Reflection, world-graph extraction, prophecy grading, final reporting, and player interaction use a higher-capability tier.

14 Discussion and limitations

These measurements describe the simulation, not real people. In particular, swarm confidence measures agreement among agents; it is not the probability that a forecast is correct. Centrality, influence, and relationship labels are also model-specific scores. They should not be read as proof of cause and effect or treated as direct measurements of human society.

Language-model responses can vary even when the numeric update rules are fixed. The chosen thresholds and weights also affect what the simulation treats as important. They should be tested with sensitivity analysis. Claims about real populations would require separate validation against real interaction data collected over time.

A Glossary

Stance	An agent's opinion on a topic axis, represented in $[-1, 1]$.
Topic	A stance axis on which the society may divide, extracted from the world prompt or prediction question.
Belief mean	Population average stance on a topic; Equation 2.
Uncertainty	Population disagreement on a topic; Equation 3.
Swarm confidence	One minus mean topic disagreement, clamped to $[0, 1]$; it measures consensus rather than predictive calibration.
Pivotal day	A day with one of the three largest positive aggregate topic drifts.
Charged interaction	A rivalry, romance, conflict, or sufficiently strong exchange that can trigger additional turns.
Public figure	An agent with influence of at least 20, whose eligible actions reach the entire population.

References

[1] J. S. Park, J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein, "Generative Agents: Interactive Simulacra of Human Behavior," in *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)*, 2023, pp. 1–22. doi: 10.1145/3586183.3606763.