WHO IS AT RISK? READING LLM DECISION BIASES INTO EVACUATION SIMULATIONS

Anonymous authors
Paper under double-blind review

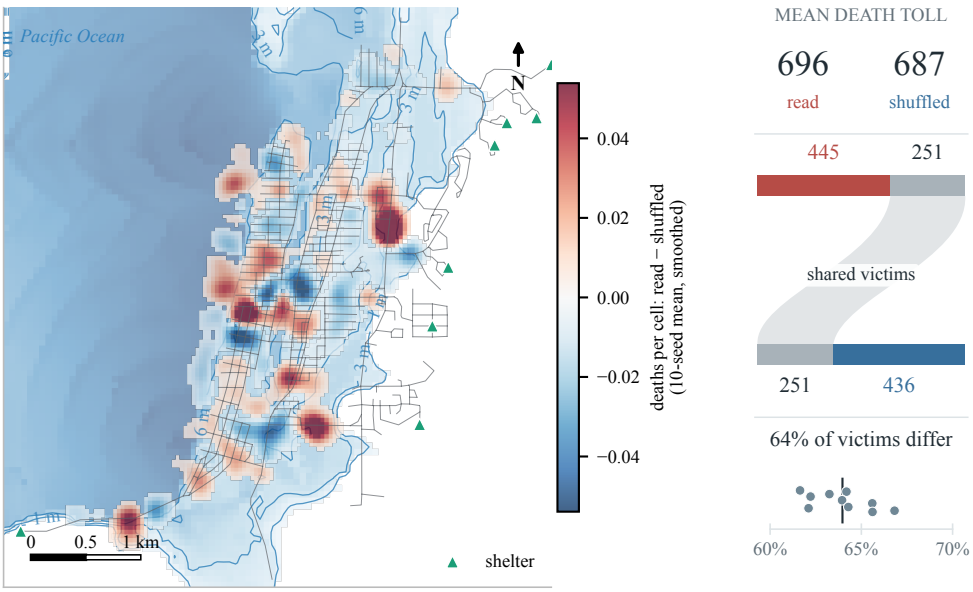

Figure 1: **Same death toll, different risk map.** A language model sets each resident's behavior. Kept by each resident (read, directly estimated from one's background) or shuffled among them, they give nearly the same death toll, yet the neighborhoods at risk and most victims change (red/blue: more deaths when read/shuffled). How can we trust who a simulation puts at risk?

ABSTRACT

In a tsunami, who survives depends on when and how each resident decides to leave. Language models now offer a way to supply these decision traits, such as how strongly a resident relies on an official warning, to every resident of an evacuation simulation, and such simulations are usually judged by whether the death toll looks plausible. This test cannot tell whether each resident received the right traits: when residents are placed at random, shuffling their traits leaves the expected death toll unchanged but changes who dies. We therefore propose checks on who holds which trait: whether shuffling changes who dies, whether the groups that fare worst match evidence from real evacuations, whether a direction read from the model does better than random directions, and whether different models agree. The checks need no data on individual behavior, and they catch errors we inject on purpose. We read six decision biases from the hidden states of four open models into two established tsunami evacuation simulators. Shuffling the traits barely changes the mean death toll but changes nearly two thirds of the victims, and models with similar death tolls differ on the checks. Warning reliance is consistent across models while herd influence is not, and all four models rate older residents as the most warning-reliant, the opposite of 2011 mortality. A plausible death toll, and even agreement between models, can hide errors in who is put at risk.

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.

1 INTRODUCTION

In a tsunami, who survives depends not only on the wave but on when each resident decides to leave: whether they heed the official warning, first gather their family, or wait until neighbors go (Lindell & Perry, 2012; Huang et al., 2016). Evacuation agent-based models (ABMs) simulate these decisions together with movement, traffic, and inundation (Mas et al., 2012; Wang et al., 2016), but their behavioral inputs are hard to obtain: surveys cover few people and few situations, so modelers fall back on distributions and hand-written rules (Widiyanto et al., 2021; Senanayake et al., 2024). Language models (LLMs) offer a new source. They already supply agents' actions or traits in social and disaster simulations (Park et al., 2023a; Aher et al., 2023; Yang et al., 2026) and can assign a decision trait to any resident profile. Prior work tests whether simulated people reproduce experimental behavior or population patterns (Aher et al., 2023; Xie et al., 2024) and finds that generated personas can carry systematic biases (Li et al., 2025a). In an evacuation, such a bias decides who is put at risk, even when the simulated death toll looks plausible.

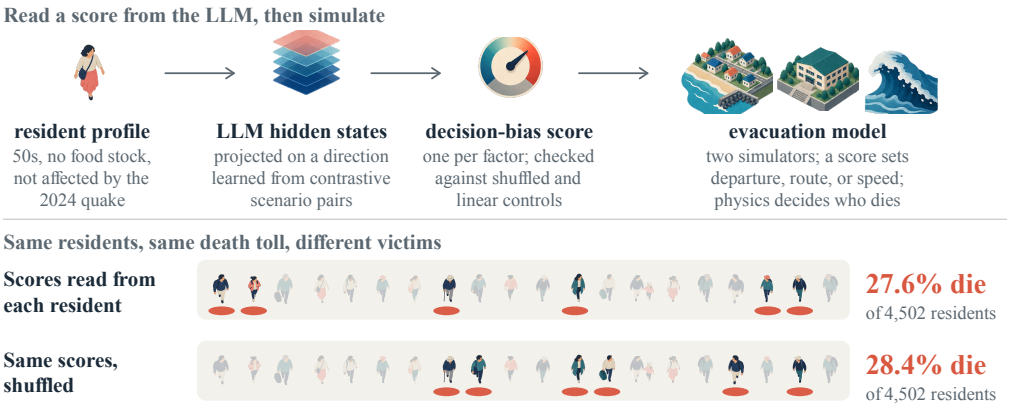

Figure 2: **Reading decision-bias scores into an evacuation simulator.** Top: a hidden-state projection gives each resident a score that sets one behavioral parameter. Bottom: twenty residents from one Seaside run, one per column (highlighted: dies); when the same scores are reassigned among the same residents, the share who die barely changes, but different residents die.

The identification problem is clearest when profiles are assigned to locations at random: permuting a fixed set of traits then leaves the *expected* aggregate unchanged (Eq. 1), although the people at risk can change (Figure 1). ABM research has long distinguished a plausible aggregate from a valid individual mechanism (Grimm et al., 2005; León-Medina, 2017; Epstein, 2023); the same distinction matters for LLM-supplied behavior (Li et al., 2025a; Zhao et al., 2026). We evaluate the assignment itself through matched simulations: first ask whether it changes individual outcomes, then compare group patterns with published evidence, hidden-state directions with random directions, independent score sources with one another, and alternative pairings of traits within agents. This uses evidence about groups without requiring a label for each simulated resident.

Our case study reads six decision biases, including reliance on official warnings and following others, from contrastive hidden-state directions (Zou et al., 2023; Kim et al., 2025). Each resident's score sets one behavioral parameter in an established tsunami evacuation simulator; the simulator still governs routes, traffic, and inundation (Wang et al., 2016; Mas et al., 2012). We test two simulators, two profile populations, four open models, and scores obtained by asking the same model directly (Figure 2; §5).

The mean death toll changes little when scores are shuffled, yet many victims change (Figure 1); aggregate outcomes also fail to distinguish sources with different evidence patterns (Table 2). Warning reliance is relatively stable across models and interfaces, while herd influence is not. All four models agree on an age ordering that runs opposite to 2011 mortality, showing why agreement alone

cannot validate a trait. Re-pairing warning and herd scores further changes who follows and who is delayed on shared roads, even though each score distribution is fixed.

We contribute a resident-level evaluation protocol for LLM-parameterized simulations and a controlled case study showing what it reveals beyond aggregate fit: consequential reassignment of risk, disagreements between score sources, a shared age ordering at odds with published evidence, and interaction between biases through a physical road network.

2 Related Work

Evacuation models and their validation. Tsunami ABMs couple departure decisions, movement, routes, and inundation (Mas et al., 2012; Wang et al., 2016; Mostafizi et al., 2017). Their behavioral differences usually come from distributions, archetypes, or survey-informed rules (Widiyanto et al., 2021; Senanayake et al., 2024; Arango & Nogal, 2026), grounded in decision models such as the protective action decision model, in which official warnings and social cues shape whether and when people leave (Lindell & Perry, 2012; Huang et al., 2016). Empirical validation of ABMs has long separated fitting aggregates from validating mechanisms (Windrum et al., 2007), and pattern-oriented modeling asks that several observed patterns be reproduced at once (Grimm et al., 2005); we use that principle to inspect which residents receive risk when a new score source is connected to an existing simulator.

Language models as simulated people. LLM agents generate believable social behavior (Park et al., 2023a), reproduce classic experiments and economic choices (Aher et al., 2023; Horton, 2023; Xie et al., 2024), and are proposed as stand-ins for survey respondents (Argyle et al., 2023; Park et al., 2024; Anthis et al., 2025); disaster simulations increasingly use their decisions (Wu et al., 2023; Li et al., 2025b; Yang et al., 2026). A parallel literature finds where they fail: their opinions track some groups more than others (Santurkar et al., 2023), they flatten and caricature identity groups (Cheng et al., 2023; Wang et al., 2025), their survey answers differ from people's in variance and in sensitivity to wording (Bisbee et al., 2024; Tjuatja et al., 2024), and generated personas carry systematic biases (Li et al., 2025a). These evaluations compare model outputs with human responses; ours asks what the assignment of traits does once it drives a physical simulation, where the evaluation target is who is at risk rather than what is said.

Reading behavioral directions. Linear probes and contrastive methods identify directions in model activations (Alain & Bengio, 2016; Hewitt & Liang, 2019; Burns et al., 2023), many concepts are represented linearly (Park et al., 2023b; Marks & Tegmark, 2023), and such directions can be read or steered (Zou et al., 2023; Li et al., 2023; Turner et al., 2023; Panickssery et al., 2023), including social perspectives (Kim et al., 2025) and character traits (Chen et al., 2025). We use directions as numerical inputs to physical ABMs, then compare the resulting resident-level assignments with direct model answers, random directions, and published evacuation patterns.

3 An Evaluation Protocol

A source q assigns each agent j a vector of trait scores $\mathbf{z}_j^q$; the simulator maps scores to behavior and returns an outcome for every agent, here whether the resident dies, and an aggregate M, here the death toll. Let p deal profiles to agents and ρ permute score vectors across profiles. When p is uniform and independent of location,

$$\mathbb{E}\left[M(\mathbf{z}_p)\right] = \mathbb{E}\left[M(\mathbf{z}_{\rho\circ p})\right] \quad \text{for any fixed permutation } \rho, \tag{1}$$

so the aggregate cannot evaluate the assignment. The shuffled control uses one fixed random permutation of intact six-score vectors across profiles: score values and correlations stay the same, but their link to profile attributes is broken. A matched read/shuffled pair starts with the same simulator seed and resident-profile allocation. The protocol replaces aggregate matching with a precondition and four validity checks, each with a stated decision rule. **C1 Reassignment (precondition):** the share of victims who survive when scores are shuffled among agents, $T = 1 - |D^q \cap D^{\text{shuf}}|/|D^q|$. *Rule:* T must exceed the share that changes when only the simulator's random stream is rerun while scores and profile assignments stay fixed; otherwise the channel does not use the scores and no validity check applies. C1 establishes that assignment matters, not that it is right. **C2 Pattern check:** the direction of group-level differences compared with published findings about real behavior (Grimm

et al., 2005); *Rule:* every group difference fixed before scoring must have the published sign, with a 95% interval over seeds that excludes zero. A failure names a trait or channel to revise; shuffled scores show whether the failing difference comes from the scores, though not whether the simulator lacks a mechanism the evidence names. **C3 Null directions:** for hidden-state sources, the same pattern statistic computed for random unit vectors at the same model layer, sampled independently of the contrastive examples; *Rule:* the weakest pre-specified group difference must exceed the 95th percentile over 1000 such directions. **C4 Convergence:** correlations between independent sources of the same trait (models, interfaces) compared with correlations across different traits, following the multitrait–multimethod logic. *Rule:* the mean same-trait correlation across sources must exceed the mean cross-trait correlation. **C5 Re-pairing:** the change in outcomes when one trait column is permuted relative to another while both distributions stay fixed, which isolates effects of how traits combine within agents. *Rule:* a source's pairing is flagged when the outcome under it lies outside the range spanned by re-pairings of the same two columns.

4 Case Study: Reading Decision Biases into Evacuation Simulators

Learning a direction for each factor. We use Qwen2.5-7B-Instruct (Qwen Team, 2024). For each factor we write pairs of short scenarios that describe the same resident, place, and moment and differ only in one sentence expressing the factor's two poles (Appendix G); for example, a resident who "treats it as an order" when the official warning is announced versus one who "assumes the official warning is probably another false alarm". Each text is followed by the same question, and we take the hidden state $\boldsymbol{h}^\ell$ at the last input token of layer ℓ. For factor k at layer ℓ, the reading direction is the first principal component of sign-randomized differences between paired hidden states (Zou et al., 2023). We orient it so that the positive pole projects higher, then score a resident j by projection:

$$\begin{aligned} \Delta_{ik}^{\ell} &= \boldsymbol{h}^{\ell}(c_{ik}^{+}) - \boldsymbol{h}^{\ell}(c_{ik}^{-}), \qquad \boldsymbol{v}_k^{\ell} = \mathrm{orient}\big(\mathrm{PC}_1\{\epsilon_i \Delta_{ik}^{\ell}\}_{i\in\mathcal{T}_k}\big)\,, \\ s_{jk} &= (\boldsymbol{v}_k^{\ell_k})^{\top} \boldsymbol{h}^{\ell_k}(c_j), \qquad z_{jk} = \frac{s_{jk} - \mu_k}{\sigma_k}. \end{aligned} \tag{2}$$

Here $\epsilon_i \in \{-1, 1\}$ is a randomly assigned sign before the differences are centered for PCA; orientation makes the mean projection of the positive-minus-negative differences positive. The selected layer is ℓ_k, while μ_k and σ_k are the mean and standard deviation of projected scores across the source profile population before assignment to residents. Thus $z_{jk} = 0$ marks that population's mean score (Figure 2, top). Equation 2 is where the LLM's hidden states enter the simulator; the ABM receives the resulting scores rather than model-generated actions. To guard against directions that merely detect wording, we choose the layer on validation wordings and report accuracy on pairs whose wordings *and* residents never appear in training. All six directions separate held-out pairs with accuracy of at least 0.84 (threshold for chance, corrected for layer selection: 0.65; Appendix F). Because contrastive scenarios state the factor explicitly, this establishes that the directions exist; Section 5 probes how the scores respond to differently worded profiles. Across profiles the six scores span about 2.9 effective dimensions (Appendix Table 10). The six factors are distinct in contrastive scenarios, but the seven survey attributes express only a few independent score dimensions.

From scores to behavior. Each factor sets exactly one behavioral parameter through a fixed function that we call its *channel* (Table 1). Every channel except panic is neutral for a typical resident ($z = 0$), who then behaves exactly as in the unmodified simulator. *Warning reliance* shifts a resident's departure time t^0, drawn from the simulator's own distribution, by $-az_W$ ($a = 12$ min on foot, 18 min by car), and a resident pushed past a threshold τ does not leave. *Herd influence* lets each such non-responder follow others with probability $\sigma(z_H)$, leaving at its own t^0. *Family priority* adds a delay of $\max(0, 100z_F)$ s; *preparedness* makes less prepared residents sometimes choose a farther shelter; *experience* scales walking or driving speed by $(1+0.1z_E)$; and *panic* pauses movement with probability $\max(0, \sigma(z_N) - 0.4)$ per step, following the common treatment of panic as disrupted movement in evacuation models (Helbing et al., 2000). Unlike the other channels it also slows typical residents; we keep it so on purpose, as a channel that can set the death toll without using the scores.

Two simulated worlds. Seaside, Oregon, is a near-field Cascadia tsunami built on the multimodal model of Wang et al. (2016): 4,502 residents walk or drive over a road network, and residents in

Table 1: Channels: each factor's score z sets one behavioral parameter. The last column is the direction in which a higher score should change mortality. t^0: the simulator's own departure time; τ: non-response threshold; σ: logistic function. Herd influence acts on the non-responders created by the warning channel, so it always runs with it.

Factor	High pole ↔ low pole	Parameter	Channel	Pred.
Warning reliance	acts on warnings ↔ skeptical	departure time	$t^0 - az_W$; stays if $> \tau$	−
Herd influence	follows others ↔ independent	response	non-responder leaves at t^0 w.p. $\sigma(z_H)$	−
Family priority	gathers family first ↔ leaves alone	added delay	$\max(0, 100z_F)$ s	+
Preparedness	prepared ↔ unprepared	shelter choice	farther shelter w.p. rising as z_P falls	−
Experience	experienced ↔ inexperienced	speed	$\times(1 + 0.1z_E)$	−
Panic	panicked ↔ calm	pause prob.	$\max(0, \sigma(z_N) - 0.4)$ per step	+

water deeper than 1 m for 2 min die. All baseline departures fall before 18 min, so $\tau = 30$ min leaves typical residents unaffected. **Arahama**, Sendai, uses TUNAMI-EVAC (Mas et al., 2012), calibrated to the 2011 Great East Japan tsunami: 2,723 pedestrians and 50 cars, departure times spread over the hour, first inundation at minute 65, and an estimated arrival time of 67 min in the model's own parameters. Half of its baseline departures fall after 30 min, so we set τ to that estimated arrival (67 min) and report the 30-minute threshold separately (§5).

Residents and controls. Each resident receives a profile made of seven attributes asked in a nationwide survey conducted two months after the 2024 Noto earthquake (Weathernews Inc., 2024a): age group, knowledge of the evacuation site at home and at work, food and mask stocks, how recently an emergency bag was checked, and relation to the 2024 Noto earthquake. We use two populations of profiles. The **attribute grid** contains every combination of these attributes (21,600 profiles), so each group is equally represented. The **survey respondents** are the survey's 19,193 respondents themselves, each carrying their own answers, so groups appear in their real proportions: 92% know their evacuation site at home, for example, against two thirds in the grid. Each profile is written as a short description and scored on all six directions. We compare these scores with two controls. **Shuffled** scores are the same intact six-score vectors reassigned by one fixed random permutation among profiles, which keeps their distribution and cross-score correlations exactly and cuts the link between a profile's attributes and its scores. We reuse this permutation in every simulator seed. A **linear rule** replaces each score with its best additive fit from the seven attributes, re-standardized; it is fitted to the LLM's scores, so it represents what a modeler could write down if they already knew the LLM's weights, and it keeps the mean and spread of the scores but not the shape of their distribution. To check that conclusions do not hinge on one set of scores, we also rerun every condition with a **second score set** for the same grid, produced by an earlier version of the reading pipeline (Appendix E). Scores are standardized within each population. Each condition runs with 10 random seeds per score source and population; a seed fixes which profile each resident receives (profiles are dealt by a seeded random permutation, in Arahama to every pedestrian and driver), so every resident can be compared across score sources.

5 What the Checks Reveal

Table 2 summarizes the case study. Shuffling any model's scores leaves its death toll nearly unchanged, and sources that fail the checks have death tolls on both sides of those that pass: the toll cannot tell them apart, the checks can. We first test the checks on known corruptions of the read warning scores (Table 3). Every corruption is flagged, none by every check, and the two most plausible ones, a score right on two of the three evidence patterns and a single readiness axis in place of warning reliance, are caught only by the null-direction check.

C1: the read scores decide who dies. As Eq. 1 requires, read and shuffled runs give nearly the same mean death toll (Appendix Figure 10), but different residents die (Figures 1 and 3). With all six channels, 51% of Seaside victims and 24% of Arahama victims change when scores are

Table 2: **Similar death tolls, different residents at risk.** Seaside attribute grid, ten matched seeds, with every channel except panic. Each row changes only the source of behavioral scores; columns report mortality, victim turnover, evidence patterns, and warning-score agreement.

Score source	Mortality change (pp)	Victims spared (%)	Evidence F / Q / K	Warning score r
Qwen2.5-7B, read	+2.3	–	✓✓✓	–
Qwen3-4B, read	+1.4	41	× ✓✓	0.77
Llama-3.2-3B, read	+2.7	47	✓✓✓	0.54
Gemma-3-4B, read	+3.8	51	× × ✓	0.58
Earlier score set	+3.5	52	✓✓×	–
Linear rule (fit to read)	+2.8	25	✓✓✓	–
Shuffled Qwen2.5-7B scores	+2.1	64	× × ×	–
Shuffled Qwen3-4B / Llama / Gemma	+1.2–+3.6	64–68	× × ×	–
Random directions (C3)	–	–	27% pass all three	–

Mortality change is in percentage points relative to no channels. Victims spared is the share of Qwen2.5-7B read-score victims who survive under the row's source (rerun noise: 2.0%). F / Q / K are food stock, Noto-quake link, and site knowledge; ✓ requires the published direction and a 95% interval excluding zero. The last column correlates warning scores with Qwen2.5-7B's. Every source fails the separate age pattern (§5).

Table 3: Controlled flaw injection. Known corruptions of the read warning scores (survey respondents); noise rows replace the stated share of variance and report rates over 20 draws. A source is flagged when it fails any check. C1 is not listed because every corruption still varies across residents.

Corruption	C2 pattern	C3 null	C4 convergence	Flagged
None (read warning scores)	pass	pass	pass	no
Sign flipped	**fail**	**fail**	**fail**	yes
Swapped with herd scores	pass	**fail**	pass	yes
Age-only score (stereotype)	**fail**	**fail**	**fail**	yes
25% noise	0% fail	100% fail	0% fail	100%
50% noise	0% fail	100% fail	0% fail	100%
75% noise	0% fail	100% fail	0% fail	100%
Pure noise	95% fail	100% fail	85% fail	100%
Stereotype right on two patterns	pass	**fail**	pass	yes
Readiness axis only (PC1)	pass	**fail**	pass	yes

shuffled, against 2.0% and 0.6% when only the simulator's random stream is rerun. Figure 1 uses five channels and shows 64% turnover. Most of this runs through the warning channel; the preparedness channel barely uses the scores, which C1 flags. The result holds when each channel constant is halved or doubled (Appendix Table 6). Read scores tie risk to circumstances: in Seaside the profile explains most of the variance in departure times with read scores and almost none with shuffled ones (Appendix Figure 11). The linear rule reproduces most of this structure but still changes 17% of victims, through interactions an additive rule omits.

C2 and C3: the patterns match evidence, beyond chance directions. Following pattern-oriented practice (Grimm et al., 2005), we compare each difference between residents with independent studies of real evacuations (Appendix Table 12). With Qwen2.5-7B's scores, three of the four point the way the evidence does, in both populations (Appendix Figure 4): residents without food stock fare worst, as less prepared residents were less likely to evacuate in 2011 (Schmöcker et al., 2023; Cabinet Office of Japan, 2012); those unaffected by the 2024 earthquake fare worse, as communities with tsunami experience had lower fatality ratios (Suppasri et al., 2016); and those who know their evacuation site fare best, as evacuation signage raised evacuation in 2011 (Schmöcker et al., 2023). The earlier score set gives a similar death toll but reverses the site-knowledge difference (Appendix E); only who dies tells the two apart. These differences arise through the warning channel, whose scenarios mention neither supplies nor evacuation sites, so their direction is what the reading supplies. Random directions in the same hidden states also produce large group differences, but in arbitrary directions: about a quarter match the evidence on all three, and the learned direction beats 95% of them on its weakest difference ($p = 0.045$).

C2 fails on age for every model. In the simulation, residents over 60 die less often than those in their fifties, whereas mortality in 2011 rose with age (Nakahara & Ichikawa, 2013). Shuffled scores

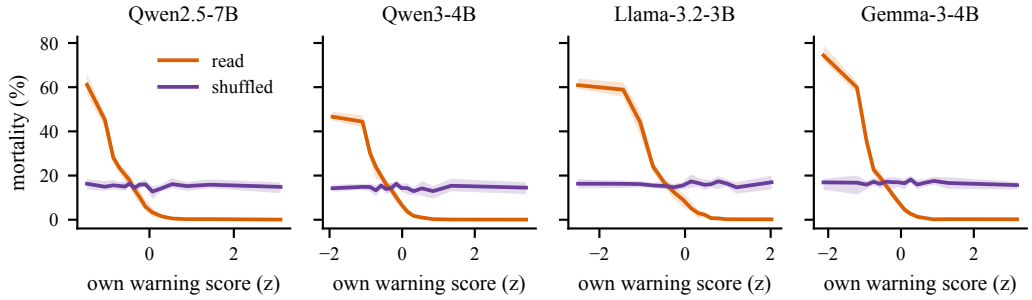

Figure 3: **Risk follows each resident's own score only when scores are read.** Mortality of Seaside residents by the warning score their own profile receives (15 bins over about 45,000 resident-seeds per curve; band: 10th–90th percentile over ten seeds), for four reading models. Read scores concentrate death on low-score residents; shuffled scores spread the same total evenly.

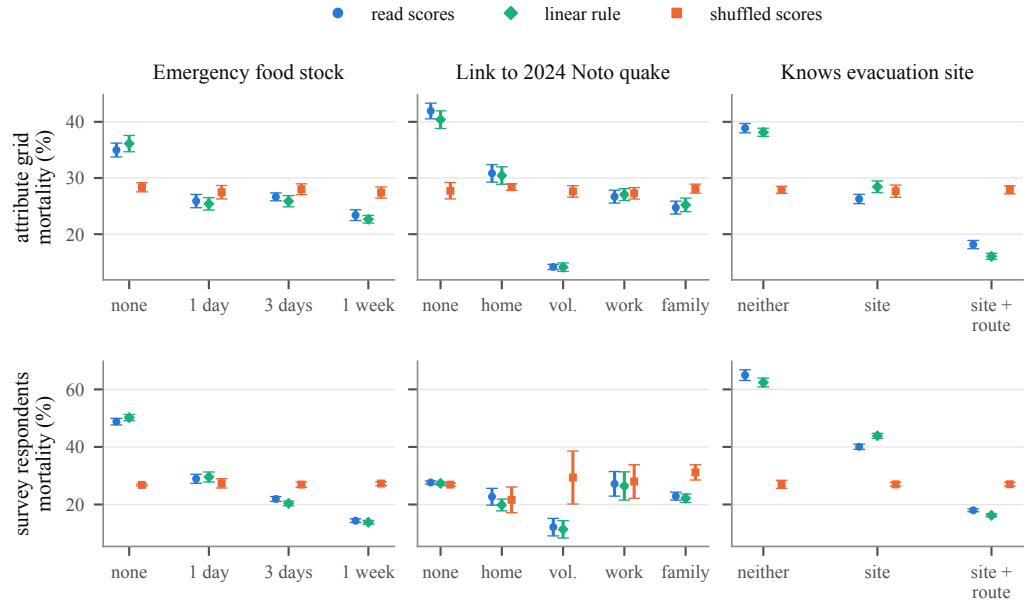

Figure 4: **Who fares worst, and does it match the evidence?** Seaside mortality with all six channels on: attribute grid (top) and survey respondents (bottom). Points are means; capped whiskers are t-based 95% intervals over ten seeds, sometimes shorter than the marker. "Vol." means volunteered after the Noto quake.

leave the two groups level, so the reversal comes from the scores: every model rates older residents as the most warning-reliant (Appendix Tables 7 and 8). Whether that ordering is wrong, or the 2011 excess reflects limits on moving that no channel represents, the check cannot say; under its rule no source passes all four patterns, and age is what to revise next. The same comparison diagnoses the channels: warning reliance and herd influence act on the behaviors the evidence ties them to, while experience, panic, and preparedness do not.

C4: warning reliance converges; herd influence does not. Llama-3.2-3B matches the same three patterns, Qwen3-4B misses food stock, and Gemma-3-4B misses food stock and reverses earthquake exposure (Table 2). Across models, warning reliance, preparedness, panic, and family priority converge, while herd influence does not (Figure 5; Appendix Table 11). The reason is a shared prior: in

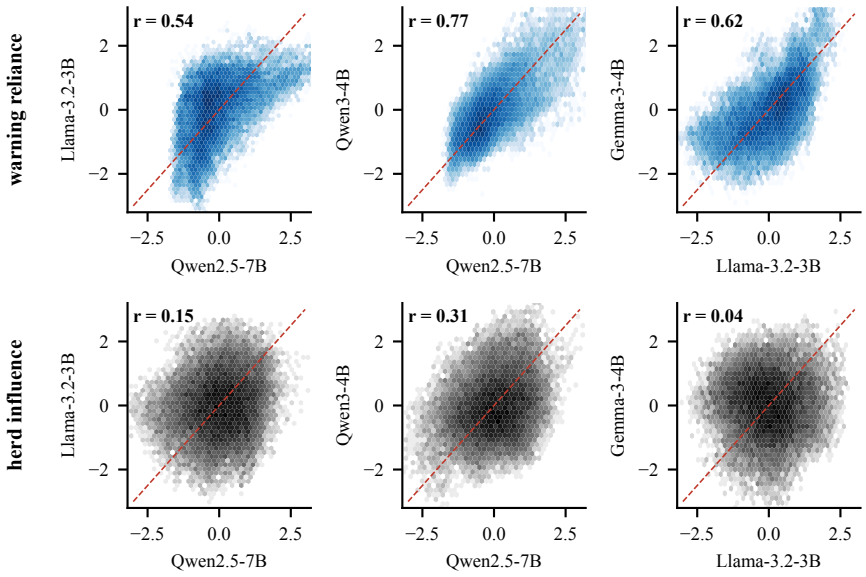

Figure 5: **Convergence, profile by profile.** The same 21,600 profiles scored by two models (density of profiles). Warning reliance lies along the diagonal for every pair of models; herd influence does not.

every model about half of the variance of the six scores (47–64%) lies on one readiness axis, and the factors that load on it most are the ones that converge. Agreement between models therefore largely reflects what they share, which is also why all four agree on the age ordering.

Reading and asking. The same model can also supply scores by answering questions (Appendix D). We describe 400 profiles in three wordings with the same facts. Greedy answers collapse to a few values, and one wording gives every resident the same delay; the expected numeric answer from the next-token distribution restores a graded score. Asked and read scores then agree on warning reliance (rank correlation above 0.9) and diverge on following others (0.25–0.49), the same trait on which the models disagree. Across the three wordings, read scores move the Seaside death toll by 1.5 points, against up to 5.1 for expected answers and 7.7 for greedy ones, so reading keeps the most structure when the same residents are described differently (Appendix Tables 13, 14, and 15).

6 C5: Re-pairing Warning and Herd through Shared Roads

Traits also combine within agents. A resident who ignores the warning follows others with probability $\sigma(z_{jH})$, so the number of followers,

$$\mathbb{E}[F \mid \mathbf{z}, \mathbf{t}^0] = \sum_j \mathbf{1}\{t_j^0 - a_j z_{jW} > \tau\}\, \sigma(z_{jH}), \tag{3}$$

depends on how warning and herd scores are paired, not only on their distributions. Comparing warning reliance alone with warning reliance plus herd influence for the same residents and seeds, the change in deaths splits exactly into followers rescued (R_F) and harmed (H_F) and non-followers rescued (R_U) and harmed (H_U):

$$D^{W+H} - D^W = H_F - R_F + H_U - R_U. \tag{4}$$

Herd influence directly saves 79.9 residents per seed, while 27.4 non-followers newly die, every one of them a driver delayed on shared roads (Figure 6). In a copy of Seaside where cars travel at a fixed speed and do not constrain one another, the direct rescue remains (80.9 per seed) and the harm to others disappears, which locates the harm in road interaction rather than in the scores. The balance depends on how many residents drive (Figure 6a): with a quarter driving, herd influence

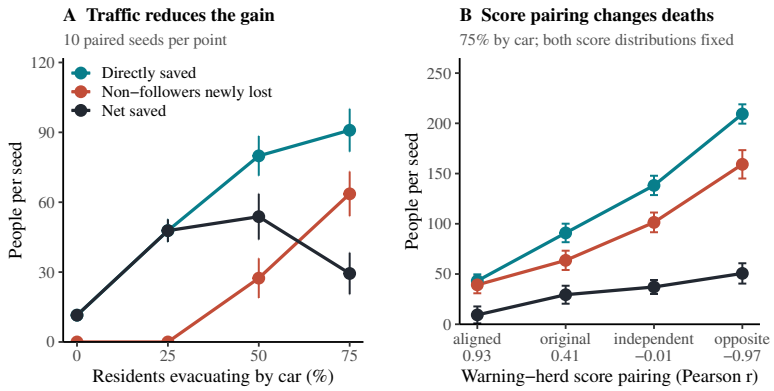

Figure 6: **Following helps some residents and delays others.** Paired Seaside runs with and without the herd channel; points average ten matched seeds, bars show 95% intervals. (a) The share of residents who drive changes the balance. (b) Re-pairing the warning and herd scores changes it with each score's distribution fixed; the original LLM scores are one of the four pairings.

harms no one, while with three quarters driving the direct gain grows but the net number saved falls, from 47.8 to 29.4 per seed. It also depends on the pairing itself: re-pairing the two score columns with both distributions fixed changes the net number saved about fivefold (Figure 6b). The original pairing lies inside that range and is not flagged. Every reading model shows the same mechanism, with magnitudes that differ because the models disagree on who is socially influenced (Appendix Table 9).

7 Discussion

Why who is at risk matters. Tsunami evacuation simulations are used to plan for events that cannot be rehearsed: where to site shelters and vertical-evacuation buildings, which routes to keep clear, where to direct warnings, and which residents need help to leave. Each of these decisions concerns particular people and places, not the total. A trait source that places risk on the wrong residents can reproduce a plausible death toll and still send help to the wrong neighborhoods and groups. The age ordering in our case study shows what is at stake: a simulation built on these scores would count older residents among the safest, while in 2011 they died at the highest rates, so a plan built on it would withdraw help from the people who most need it.

Impact. Language models make it cheap to give every simulated resident an individual behavior, which evacuation models have long lacked, and they are being adopted for this purpose. The benefit is real only if the right behavior reaches the right people. Our checks let modelers and emergency managers test this before a simulation informs a plan, using only the simulator and the group-level findings they already rely on; they establish consistency with evidence rather than individual accuracy, and they show which traits can be trusted, such as warning reliance here, and which should be treated as uncertain and varied across scenarios, such as herd influence.

What the language model adds. A rule that adds up attribute weights fitted to the read scores passes the checks as well, but it cannot be written without them. Published studies give the direction of a few attribute effects, not how much each of seven attributes shifts each of six traits; the language model supplies those magnitudes for any profile, and the fitted rule makes them explicit and editable.

8 Conclusion

Language-model-parameterized evacuation simulations can match the death toll while placing risk on the wrong residents. Checking who holds which trait exposes this without individual ground truth, separates sources that the death toll cannot, and shows that agreement between models, which here includes an age ordering the evidence contradicts, is not validity.

Ethics Statement

Resident profiles use seven categorical answers from the 19,193-response CSV of Weathernews's 2024 disaster-preparedness survey, available from the publisher's public data page (Weathernews Inc., 2024a). Weathernews's release invites use of these data for disaster research and risk-reduction activities (Weathernews Inc., 2024b); we rely on that stated purpose for our analysis. The CSV contains no names or addresses, and our profiles omit its prefecture fields. We report respondent results in aggregate and direct readers to the publisher's download rather than republishing the row-level CSV. Evidence about the 2011 and other events comes from published studies. The simulations assign risk to residents by profile, and the paper shows that language-model scores can carry orderings that published evidence does not support, such as rating older residents as the most warning-reliant; results of this kind should not inform evacuation planning for real groups without behavioral validation.

Reproducibility Statement

Appendix H lists the scripts that run every simulation and generate every reported number, table, and figure; each run writes one record per resident with its profile, scores, seed, and channel configuration. Direction learning, layer selection, and scenario pairs are described in Section 4 and Appendices F and G; channel definitions and constants are in Table 1 and Appendix Table 6. Code and grid-profile score files will be released with the paper; the respondent-level survey CSV is available from Weathernews (Weathernews Inc., 2024a).

AI Use Statement

AI in the experiments. Four open language models supplied resident-level decision-bias scores through hidden-state projections; we also obtained scores by asking one model directly. The scores set behavioral parameters in two existing tsunami evacuation simulators. Simulation outcomes and evaluation statistics were computed by code, not judged by a language model.

AI in research and writing. We used generative AI assistants to help write analysis and plotting code, search related work, interpret results, prepare figures, and edit the manuscript. The authors reviewed AI-assisted text and artifacts and take responsibility for the final paper and results.

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*, volume 202 of *Proceedings of Machine Learning Research*, pp. 337–371, 2023. URL https://proceedings.mlr.press/v202/aher23a.html.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Jacy Reese Anthis, Ryan Liu, Sean M. Richardson, Austin C. Kozlowski, Bernard Koch, James Evans, Erik Brynjolfsson, and Michael Bernstein. Position: LLM social simulations are a promising research method. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

Erica Arango and Maria Nogal. Flood emergency mobility: Archetype-based modeling of human behavior. In *Proceedings of the 23rd International ISCRAM Conference*, volume 23, 2026.

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351, 2023.

James Bisbee, Joshua D. Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer M. Larson. Synthetic replacements for human survey data? the perils of large language models. *Political Analysis*, 32 (4), 2024. doi: 10.1017/pan.2024.5.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *International Conference on Learning Representations*, 2023.

Cabinet Office of Japan. Resident questionnaire survey on earthquake and tsunami evacuation during the Great East Japan Earthquake: Main results. Technical report, Cabinet Office, Government of Japan, 2012. URL https://www.bousai.go.jp/jishin/tsunami/hinan/pdf/20121221_chousa1_1.pdf. Face-to-face drop-off survey of 11,400 residents of 27 municipalities in Iwate, Miyagi, and Fukushima. In Japanese.

Runjin Chen, Andy Arditi, Henry Sleight, Owain Evans, and Jack Lindsey. Persona vectors: Monitoring and controlling character traits in language models. *arXiv preprint arXiv:2507.21509*, 2025.

Myra Cheng, Tiziano Piccardi, and Diyi Yang. CoMPosT: Characterizing and evaluating caricature in LLM simulations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

Julie L. Demuth, Rebecca E. Morss, Jeffrey K. Lazo, and Craig W. Trumbo. The effects of past hurricane experiences on evacuation intentions through risk perception and efficacy beliefs: A mediation analysis. *Weather, Climate, and Society*, 8(4):327–344, 2016. doi: 10.1175/WCAS-D-15-0074.1.

John Drury, Chris Cocking, and Steve Reicher. Everyone for themselves? a comparative study of crowd solidarity among emergency survivors. *British Journal of Social Psychology*, 48(3):487–506, 2009. doi: 10.1348/014466608X357893.

Joshua M. Epstein. Inverse generative social science: Backward to the future. *Journal of Artificial Societies and Social Simulation*, 26(2):9, 2023.

Volker Grimm, Eloy Revilla, Uta Berger, Florian Jeltsch, Wolf M. Mooij, Steven F. Railsback, Hans-Hermann Thulke, Jacob Weiner, Thorsten Wiegand, and Donald L. DeAngelis. Pattern-oriented modeling of agent-based complex systems: Lessons from ecology. *Science*, 310(5750):987–991, 2005.

Dirk Helbing, Illés Farkas, and Tamás Vicsek. Simulating dynamical features of escape panic. *Nature*, 407:487–490, 2000.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

John J. Horton. Large language models as simulated economic agents: What can we learn from homo silicus? Working Paper 31122, National Bureau of Economic Research, 2023.

Shih-Kai Huang, Michael K. Lindell, and Carla S. Prater. Who leaves and who stays? a review and statistical meta-analysis of hurricane evacuation studies. *Environment and Behavior*, 48(8):991–1029, 2016. doi: 10.1177/0013916515578485.

Junsol Kim, James Evans, and Aaron Schein. Linear representations of political perspective emerge in large language models. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/1557cc3e1a67f985dbdb162db9963dd0-Abstract-Conference.html.

Francisco J. León-Medina. Analytical sociology and agent-based modeling: Is generative sufficiency sufficient? *Sociological Theory*, 35(3):157–178, 2017.

Ang Li, Haozhe Chen, Hongseok Namkoong, and Tianyi Peng. LLM generated persona is a promise with a catch. In *Advances in Neural Information Processing Systems*, volume 38, 2025a. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/5fc150b75e8b509735564de16bc3527b-Abstract-Position_Paper_Track.html. Position Paper Track.

Bo Li, Junwei Ma, Kai Yin, Yiming Xiao, Chia-Wei Hsu, and Ali Mostafavi. Disaster management in the era of agentic ai systems: a vision for collective human-machine intelligence for augmented resilience. *arXiv preprint arXiv:2510.16034*, 2025b.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Michael K. Lindell and Ronald W. Perry. The protective action decision model: Theoretical modifications and additional evidence. *Risk Analysis*, 32(4):616–632, 2012.

Fumiyasu Makinoshima, Yusuke Oishi, Masaharu Nakagawa, Shosuke Sato, and Fumihiko Imamura. Revealing complex tsunami evacuation process patterns induced by social interactions: A case study in Ishinomaki. *International Journal of Disaster Risk Reduction*, 58:102182, 2021. doi: 10.1016/j.ijdrr.2021.102182.

Fumiyasu Makinoshima, Saki Yotsui, Shosuke Sato, and Fumihiko Imamura. Massive geolocation data reveal evacuation behaviour during the 2024 Noto Peninsula earthquake and tsunami. *arXiv preprint arXiv:2412.05795*, 2024.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:2310.06824*, 2023.

Erick Mas, Anawat Suppasri, Fumihiko Imamura, and Shunichi Koshimura. Agent-based simulation of the 2011 great east japan earthquake/tsunami evacuation: An integrated model of tsunami inundation and evacuation. *Journal of Natural Disaster Science*, 34(1):41–57, 2012. doi: 10.2328/jnds.34.41.

Takahito Mikami, Tomoya Shibayama, Miguel Esteban, and Rafael Aránguiz. Comparative analysis of triggers for evacuation during recent tsunami events. *Natural Hazards Review*, 21(3):04020022, 2020. doi: 10.1061/(ASCE)NH.1527-6996.0000386.

Alireza Mostafizi, Haizhong Wang, Dan Cox, Lori A. Cramer, and Shangjia Dong. Agent-based tsunami evacuation modeling of unplanned network disruptions for evidence-driven resource allocation and retrofitting strategies. *Natural Hazards*, 88(3):1347–1372, 2017.

Shinji Nakahara and Masao Ichikawa. Mortality in the 2011 tsunami in Japan. *Journal of Epidemiology*, 23(1):70–73, 2013. doi: 10.2188/jea.JE20120114.

Nina Panickssery, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering llama 2 via contrastive activation addition. *arXiv preprint arXiv:2312.06681*, 2023.

Joon Sung Park, Joseph O'Brien, Carrie Jun 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*. ACM, 2023a.

Joon Sung Park, Carolyn Q. Zou, Jonne Kamphorst, Niles Egan, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Percy Liang, Robb Willer, and Michael S. Bernstein. LLM agents grounded in self-reports enable general-purpose simulation of individuals. *arXiv preprint arXiv:2411.10109*, 2024.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. *arXiv preprint arXiv:2311.03658*, 2023b.

Enrico L. Quarantelli. The nature and conditions of panic. *American Journal of Sociology*, 60(3): 267–275, 1954. doi: 10.1086/221536.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, 2023.

Jan Dirk Schmöcker, Jun Ji, Fajar Prawira Belgiawan, and Nobuhiro Uno. Evacuation decisions during the great east japan earthquake. *Findings*, 2023.

Gayani P. D. P. Senanayake, Minh Kieu, Yang Zou, and Kim Dirks. Agent-based simulation for pedestrian evacuation: A systematic literature review. *International Journal of Disaster Risk Reduction*, 111:104705, 2024.

Anawat Suppasri, Natsuki Hasegawa, Fumiyasu Makinoshima, Fumihiko Imamura, Panon Latcharote, and Simon Day. An analysis of fatality ratios and the factors that affected human fatalities in the 2011 Great East Japan tsunami. *Frontiers in Built Environment*, 2:32, 2016. doi: 10.3389/fbuil.2016.00032.

Lindia Tjuatja, Valerie Chen, Sherry Tongshuang Wu, Ameet Talwalkar, and Graham Neubig. Do LLMs exhibit human-like response biases? a case study in survey design. *Transactions of the Association for Computational Linguistics*, 12, 2024.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023.

Angelina Wang, Jamie Morgenstern, and John P. Dickerson. Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence*, 7, 2025. doi: 10.1038/s42256-025-00986-z.

Haizhong Wang, Alireza Mostafizi, Lori A. Cramer, Dan Cox, and Hyoungsu Park. An agent-based model of a multimodal near-field tsunami evacuation: Decision-making and life safety. *Transportation Research Part C: Emerging Technologies*, 64:86–100, 2016.

Weathernews Inc. Gensai chōsa 2024: Survey data. `https://weathernews.jp/quake_tsunami_311/data2024.html`, 2024a. Public CSV of 19,193 survey responses collected 28 February–3 March 2024. In Japanese.

Weathernews Inc. Gensai chōsa 2024: Survey results and data release. `https://jp.weathernews.com/news/46285`, 2024b. Press release, 14 March 2024. In Japanese.

Sinung Widiyanto, Dimas Adi, Nadhila Nurdin, et al. Agent-based simulation for evaluating the effect of different walking and driving speed on disaster evacuation in aceh. *IPTEK the Journal of Engineering*, 7(2):50–58, 2021.

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

Zengqing Wu, Run Peng, Xu Han, Shuyuan Zheng, Yixin Zhang, and Chuan Xiao. Smart agent-based modeling: On the use of large language models in computer simulations. *arXiv preprint arXiv:2311.06330*, 2023.

Chengxing Xie, Canyu Chen, Feiran Jia, Ziyu Ye, Shiyang Lai, Kai Shu, Jindong Gu, Adel Bibi, Ziniu Hu, David Jurgens, James Evans, Philip H. S. Torr, Bernard Ghanem, and Guohao Li. Can large language model agents simulate human trust behavior? In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL `https://proceedings.neurips.cc/paper_files/paper/2024/file/1cb57fcf7ff3f6d37eebae5becc9ea6d-Paper-Conference.pdf`.

Sen Yang, Luis Ceferino, Yi Zhang, Chen Gu, Tong Guo, and Gen Kondo. When agents learn to think: Large language model-enhanced agent-based modeling for crowd evacuation in disaster scenarios. *Reliability Engineering & System Safety*, 269:112056, 2026.

Nam Yi Yun and Masanori Hamada. Evacuation behavior and fatality rate during the 2011 Tohoku-Oki earthquake and tsunami. *Earthquake Spectra*, 31(3):1237–1265, 2015. doi: 10.1193/082013EQS234M.

Peter Zhao, Dan Hoang Pham, and Nicholas Vincent. Mechanism plausibility in generative agent-based modeling. In *Proceedings of the 2026 ACM Conference on Fairness, Accountability, and Transparency*, pp. 5915–5940, 2026.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A Simulation Results

Tables 4 and 5 give each channel's effect with read, linear-rule, and shuffled scores in the two simulators, and Figure 9 shows where each channel enters a resident's decision. Figure 10 shows that read and shuffled scores give the same death toll in every condition; Figures 11 and 12 show that they differ in who departs when; Figure 13 repeats the group comparison for Arahama. Table 6 varies each channel constant. Figure 7 shows how the Seaside tsunami floods the town, and Figure 8 where residents die in one seed and how: residents who never leave die in low-lying neighborhoods, pedestrians along the routes to high ground, and drivers in a queue at one crossing, the congestion behind Section 6.

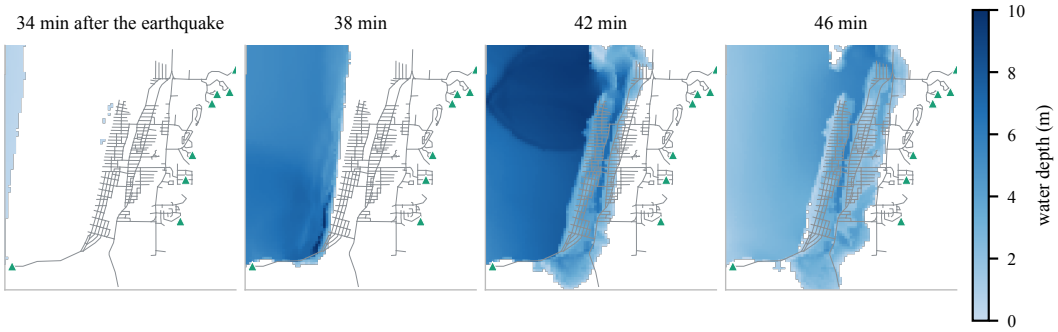

Figure 7: The Seaside tsunami: water depth at four times after the earthquake, over the road network (grey) and evacuation shelters (green triangles). The wave reaches the shore at about 35 minutes and floods most of the town within ten minutes.

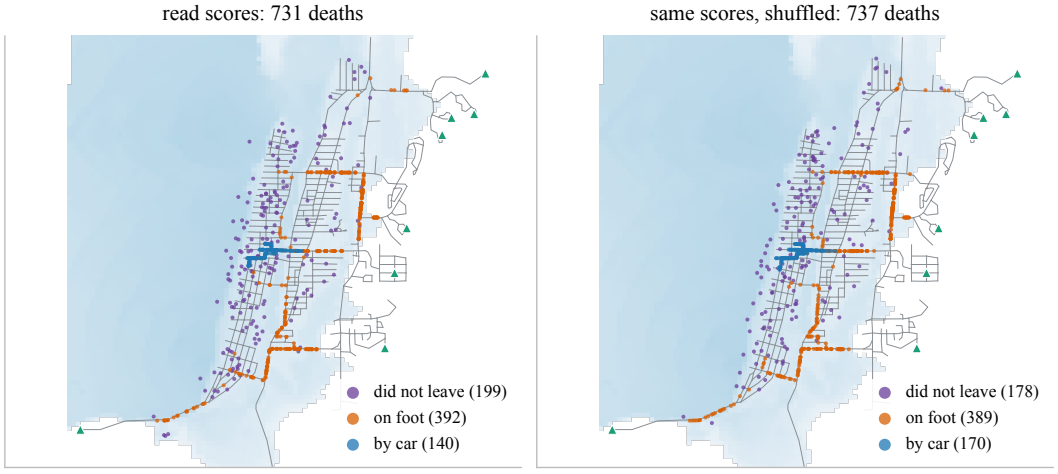

Figure 8: Where residents die in one Seaside run (seed 1, all channels except panic), with Qwen2.5-7B scores read for each resident (left) or shuffled among the same residents (right); shading: maximum inundation. Each point is a victim's final position, colored by how they died. The totals are nearly equal, but the mix differs: read scores leave more residents who never evacuate and fewer who die in traffic.

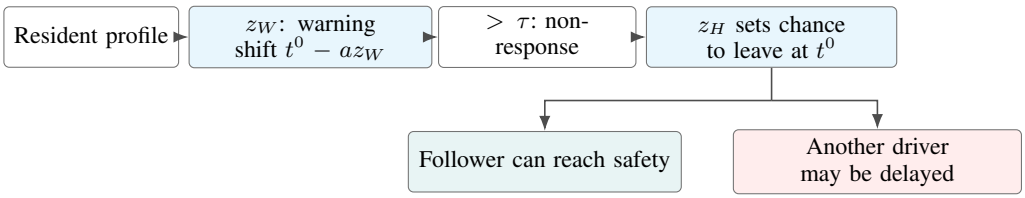

Figure 9: **One decision, two outcomes.** Warning reliance determines who is eligible to respond to the assumed social cue. Herd influence determines whether each non-responder leaves. Movement can then save the follower and delay a resident whose own decision did not change.

Table 4: Seaside: paired differences from the matched baseline (mortality 13.16%, SD 0.70, over 10 seeds). ΔMort. in percentage points with 95% CI across seeds; ΔDep. in seconds (non-responders censored at 60 min); Non-resp.: share of residents who never leave; R^2: share of agent-level variance in departure time and in death (a 0/1 outcome, linear probability model) explained by the resident profile attributes.

Condition	Scores	ΔMort.	95% CI	ΔDep.	Non-resp. (%)	R^2_{dep}	R^2_{death}
Warning reliance	Read	+1.20	[+0.61, +1.79]	+245	6.3	0.71	0.27
	Linear	+3.65	[+3.26, +4.04]	+259	11.2	0.77	0.36
	Shuffled	+1.24	[+0.72, +1.77]	+245	6.3	0.01	0.01
Herd (+ warning)	Read	+0.02	[−0.46, +0.50]	+188	4.3	0.65	0.20
	Linear	+0.85	[+0.45, +1.25]	+143	7.2	0.67	0.25
	Shuffled	−0.03	[−0.50, +0.44]	+186	4.2	0.00	0.00
Family priority	Read	+0.71	[+0.36, +1.05]	+41	0.0	0.27	0.01
	Linear	+0.56	[+0.37, +0.75]	+40	0.0	0.30	0.01
	Shuffled	+0.64	[+0.48, +0.80]	+40	0.0	0.01	0.01
Preparedness	Read	−0.18	[−0.43, +0.07]	−0	0.0	0.01	0.01
	Linear	−0.11	[−0.35, +0.13]	+0	0.0	0.01	0.01
	Shuffled	−0.15	[−0.36, +0.07]	+0	0.0	0.01	0.01
Experience	Read	+1.49	[+1.24, +1.75]	+0	0.0	0.01	0.01
	Linear	+1.57	[+1.23, +1.90]	−0	0.0	0.01	0.01
	Shuffled	+1.51	[+1.21, +1.80]	+0	0.0	0.01	0.01
Panic	Read	+13.85	[+13.54, +14.16]	−0	0.0	0.01	0.04
	Linear	+14.00	[+13.70, +14.29]	−0	0.0	0.01	0.04
	Shuffled	+13.94	[+13.65, +14.24]	+0	0.0	0.01	0.01
All six	Read	+14.53	[+14.07, +15.00]	+229	4.3	0.66	0.23
	Linear	+14.31	[+13.81, +14.81]	+181	7.1	0.67	0.26
	Shuffled	+14.65	[+14.21, +15.10]	+229	4.3	0.00	0.00

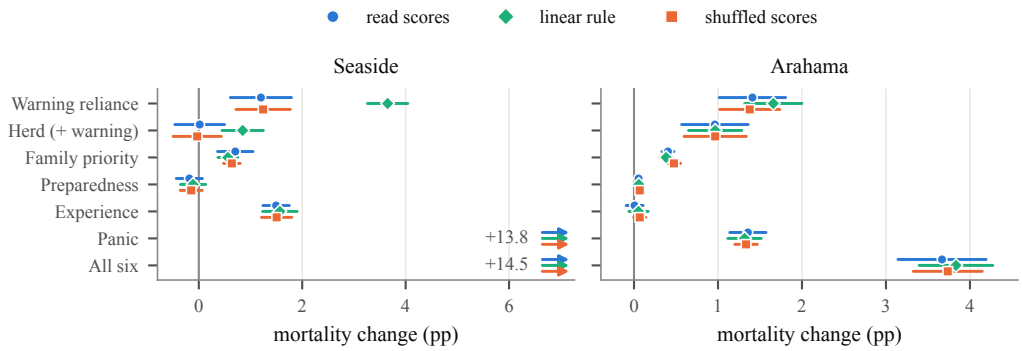

Figure 10: Mortality change when a channel is switched on, against the same simulation with every channel off (95% intervals, 10 seeds). Read and shuffled scores give the same death toll in every condition; the linear rule departs only where its score distribution differs (warning and herd in Seaside).

Table 5: Arahama (non-response threshold at tsunami arrival, 67 min): paired differences from the matched baseline (mortality 18.68%, SD 0.42, over 10 seeds). ΔMort. in percentage points with 95% CI across seeds; ΔDep. in seconds (non-responders censored at 60 min); Non-resp.: share of residents who never leave; R^2: share of agent-level variance in departure time and in death (a 0/1 outcome, linear probability model) explained by the resident profile attributes.

Condition	Scores	ΔMort.	95% CI	ΔDep.	Non-resp. (%)	R^2_{dep}	R^2_{death}
Warning reliance	Read	+1.41	[+1.01, +1.81]	+23	15.0	0.16	0.04
	Linear	+1.66	[+1.32, +2.00]	+8	15.2	0.19	0.05
	Shuffled	+1.38	[+1.02, +1.73]	+24	15.1	0.01	0.01
Herd (+ warning)	Read	+0.96	[+0.57, +1.36]	+22	7.9	0.16	0.03
	Linear	+0.97	[+0.65, +1.28]	+6	8.0	0.19	0.04
	Shuffled	+0.97	[+0.60, +1.34]	+23	7.8	0.01	0.01
Family priority	Read	+0.40	[+0.33, +0.48]	+33	0.0	0.01	0.01
	Linear	+0.38	[+0.31, +0.45]	+32	0.0	0.01	0.01
	Shuffled	+0.48	[+0.40, +0.56]	+33	0.0	0.01	0.01
Preparedness	Read	+0.05	[+0.03, +0.08]	+0	0.0	0.01	0.01
	Linear	+0.06	[+0.02, +0.09]	+0	0.0	0.01	0.01
	Shuffled	+0.07	[+0.04, +0.09]	+0	0.0	0.01	0.01
Experience	Read	+0.01	[−0.09, +0.11]	+0	0.0	0.01	0.01
	Linear	+0.05	[−0.06, +0.17]	+0	0.0	0.01	0.01
	Shuffled	+0.07	[−0.00, +0.14]	+0	0.0	0.01	0.01
Panic	Read	+1.36	[+1.14, +1.57]	+0	0.0	0.01	0.01
	Linear	+1.32	[+1.12, +1.52]	+0	0.0	0.01	0.01
	Shuffled	+1.33	[+1.20, +1.47]	+0	0.0	0.01	0.01
All six	Read	+3.67	[+3.14, +4.19]	+54	7.9	0.16	0.05
	Linear	+3.83	[+3.39, +4.27]	+37	8.0	0.19	0.07
	Shuffled	+3.74	[+3.33, +4.15]	+56	7.8	0.01	0.01

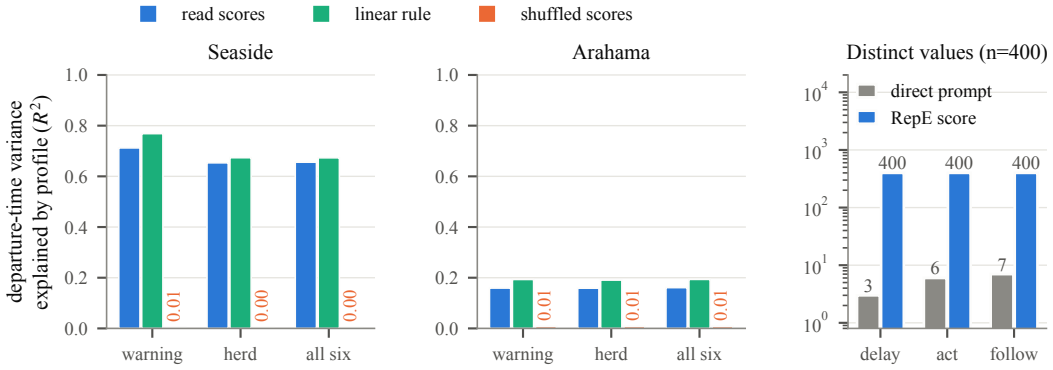

Figure 11: Left and middle: share of the variance in residents' departure times explained by their profiles, for read, linear-rule, and shuffled scores. Right: distinct values across the same 400 profiles when the model is asked directly (most under any one wording) and when scores are read.

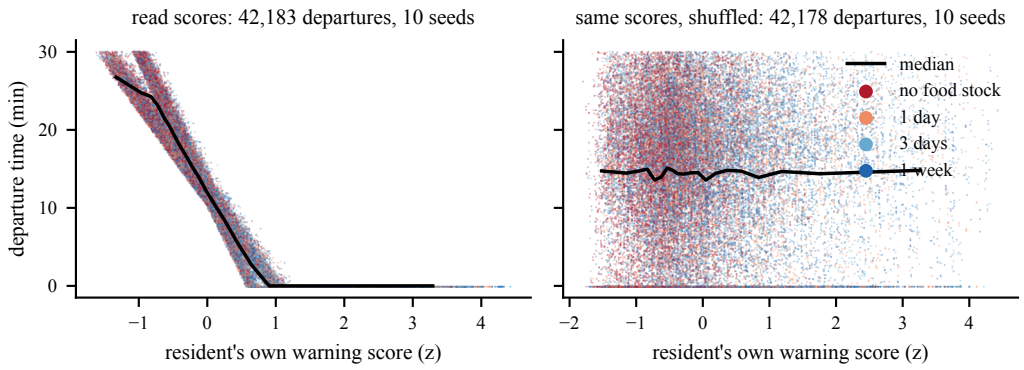

Figure 12: Every Seaside resident's departure time against the warning score of its own profile, ten seeds (read left, shuffled right); color: emergency food stock; line: median by score bin.

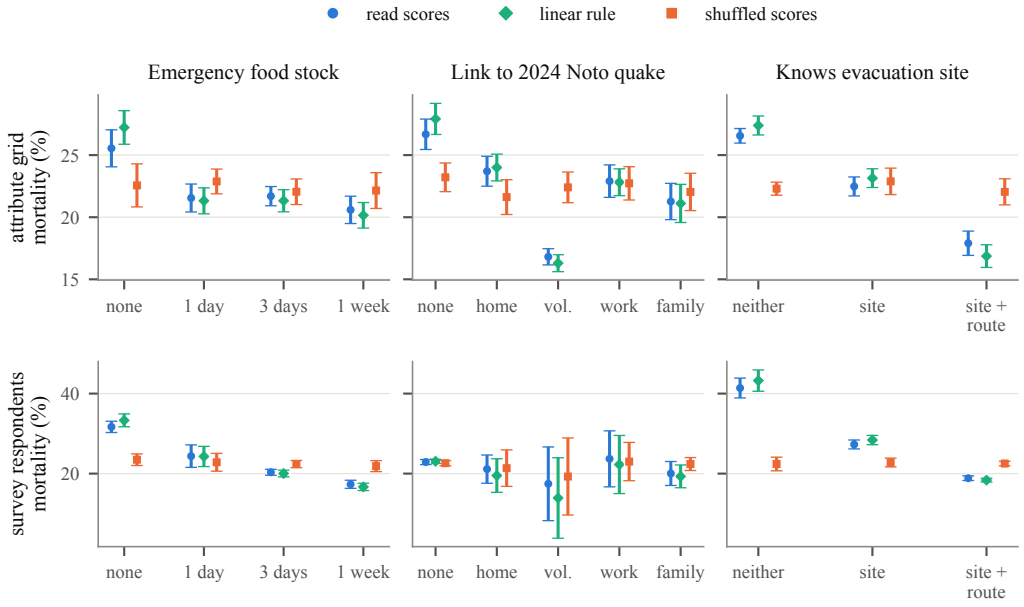

Figure 13: Arahama (threshold at arrival): mortality by profile attribute with all six channels on, 95% intervals across seeds, for the attribute grid (top) and the survey respondents (bottom).

Table 6: Sensitivity to channel constants: Seaside, all channels except panic, attribute grid, read versus shuffled scores over ten matched seeds. Read−shuffled: seed-paired mortality difference (pp, 95% interval). Victims: % of read-score victims who survive under shuffled scores. Group columns: mortality (%) without / with food stock, without / with 2024 earthquake exposure, without / with site knowledge, under read scores.

Variant	Mort. (%)	Read−shuffled	Victims	Food	Quake	Knowledge
Default	15.5	+0.20 [-0.24, +0.64]	64	21.3 / 13.5	35.1 / 10.5	22.5 / 12.0
Warning shift $\times 0.5$	9.8	+0.19 [-0.12, +0.51]	51	12.5 / 9.0	14.2 / 8.8	15.0 / 7.3
Warning shift $\times 2$	23.1	+0.04 [-0.44, +0.53]	68	31.2 / 20.5	51.7 / 15.9	29.5 / 20.0
Threshold $\tau = 20$ min	21.2	+0.28 [-0.13, +0.70]	71	28.5 / 18.8	51.4 / 13.5	26.3 / 18.7
Threshold $\tau = 40$ min	13.4	+0.19 [-0.05, +0.43]	58	18.2 / 11.8	24.8 / 10.5	20.5 / 9.9
Family delay $\times 0.5$	15.1	+0.24 [-0.17, +0.66]	65	20.9 / 13.2	35.2 / 10.0	21.8 / 11.8
Family delay $\times 2$	16.2	+0.17 [-0.29, +0.63]	63	21.8 / 14.3	34.9 / 11.4	24.1 / 12.3
Speed scaling $\times 0.5$	15.0	+0.38 [+0.01, +0.76]	64	21.1 / 13.0	35.5 / 9.8	21.0 / 12.1
Speed scaling $\times 2$	16.3	+0.09 [-0.41, +0.58]	65	21.9 / 14.4	34.3 / 11.7	24.8 / 12.1

B Models and Convergence

Table 7 compares the four reading models on the attribute grid and Table 8 on the survey respondents; Table 9 repeats the herd mechanism of Section 6 for each model. Table 10 gives the correlations among one model's six scores, and Table 11 the convergence of each factor across models.

Table 7: The same tests across reading models (attribute grid). Readable: factors above the layer-corrected chance threshold. ρ: rank correlation across wordings. Cues: factors whose high-pole cue outranks the low-pole cue in $\geq$80% of profiles. $|\Delta|$: largest read-minus-shuffled mortality difference (pp). Shuffled, vs. Qwen2.5: % of victims that differ. Evidence: higher mortality without food stock, without quake exposure, and without site knowledge (all channels except panic; the 95% interval over seeds must exclude zero, as in Table 2), and at 60+ (all six channels, sign only).

					Shuffled			vs. Qwen2.5	
Model	Readable	ρ	Cues	$\lvert\Delta\rvert$	Sea.	Ara.	Evidence (F Q K A)	Sea.	Ara.
Qwen2.5-7B	6/6	0.71	5/6	0.12	51	24	✓ ✓ ✓ ×	–	–
Qwen3-4B	6/6	0.54	4/6	0.32	46	23	× ✓ ✓ ×	29	14
Llama-3.2-3B	6/6	0.72	6/6	0.28	55	30	✓ ✓ ✓ ×	28	18
Gemma-3-4B	5/6	0.54	4/6	0.14	52	26	× × ✓ ×	36	17

Table 8: The simulation tests of Table 7 for the survey-respondent population. ΔMort.: mortality change with all six channels on (pp). $|\Delta|$: largest seed-paired mean difference between read and shuffled scores over the seven conditions (pp). Shuffled: % of victims (all six channels) who survive when scores are shuffled. Evidence as in Table 7 (Seaside).

	ΔMort.		$\lvert\Delta\rvert$		Shuffled		
Model	Sea.	Ara.	Sea.	Ara.	Sea.	Ara.	Evidence (F Q K A)
Qwen2.5-7B	+14.3	+4.1	0.32	0.28	54	27	✓ ✓ ✓ ×
Qwen3-4B	+14.2	+3.9	0.24	0.11	51	26	✓ ✓ ✓ ×
Llama-3.2-3B	+11.6	+4.2	0.39	0.34	60	32	✓ ✓ ✓ ×
Gemma-3-4B	+14.8	+4.1	0.19	0.26	53	27	✓ ✓ ✓ ×

Table 9: The direct gain and road cost persist across four LLM score sources in Seaside with 75% driving. Entries are means over ten paired seeds; car delay is the mean change in arrival time for non-following drivers safe in both runs. Model-to-model counts are descriptive because the score distributions also differ.

Score source	Followers	Saved	Newly lost	Net saved	Car delay (min)
Qwen2.5-7B	131.0	90.9	63.6	29.4	0.45
Gemma-3-4B	136.7	91.3	73.4	18.5	0.53
Llama-3.2-3B	271.7	178.0	122.4	55.6	0.93
Qwen3-4B	263.0	169.2	125.3	45.2	0.96

Table 10: Correlation of the six read scores across profiles: attribute grid below the diagonal, survey respondents above. Effective number of dimensions (participation ratio of the eigenvalues): 2.9 (grid), 2.1 (respondents).

	Warning	Herd	Family	Prepared.	Experience	Panic
Warning		0.86	0.09	0.93	0.58	-0.74
Herd	0.41		0.43	0.77	0.35	-0.49
Family	0.15	0.58		-0.07	-0.32	0.43
Prepared.	0.86	0.24	-0.15		0.71	-0.89
Experience	0.24	-0.22	-0.41	0.49		-0.73
Panic	-0.43	0.28	0.57	-0.70	-0.64	

Table 11: Cross-source convergence of the read scores over the attribute grid. Same factor: correlation between two reading models' scores for the same factor, mean and range over the six model pairs (Qwen2.5-7B, Qwen3-4B, Llama-3.2-3B, Gemma-3-4B). Other factors: mean absolute correlation between one model's score for this factor and another model's score for a different factor. A factor converges when the first exceeds the second.

Factor	Same factor	Range	Other factors
Warning reliance	+0.65	[+0.54, +0.77]	0.37
Preparedness	+0.62	[+0.43, +0.80]	0.42
Panic	+0.63	[+0.47, +0.81]	0.41
Family priority	+0.44	[+0.29, +0.56]	0.30
Experience	+0.35	[+0.05, +0.61]	0.38
Herd influence	+0.10	[-0.11, +0.31]	0.17

C Evidence for Each Channel

Table 12 compares each channel with independent studies of real evacuations; it underlies the pattern check and the channel diagnosis in Section 5.

Table 12: Channels compared with independent human studies. GEJE: 2011 Great East Japan earthquake and tsunami.

Channel	Independent human evidence	Our simulation	Verdict
Warning reliance	Official warnings have consistently significant effects across 38 studies of real hurricane evacuations (Huang et al., 2016); in GEJE, not knowing the warning had been issued was one of the two most common reasons for staying (Cabinet Office of Japan, 2012), and departure time is among the strongest predictors of survival for 1,153 GEJE witnesses (Yun & Hamada, 2015).	Warning-reliant residents leave earlier and die less; skeptics may not leave.	Consistent
Herd influence	Observed social cues have consistently significant effects (Huang et al., 2016); in GEJE, calls from others triggered about as many evacuations as the warning (Cabinet Office of Japan, 2012; Schmöcker et al., 2023), and sighted movements triggered further evacuation (Makinoshima et al., 2021).	Followers leave and survive.	Consistent
Family priority	Over 70% of GEJE residents tried to confirm family safety and about half went to check on or collect others before the wave, delaying departure (Cabinet Office of Japan, 2012), yet reunification also moved families to high ground (Makinoshima et al., 2021; 2024).	The channel only delays.	Partial
Experience	Experience changes evacuation *intentions* through risk perception, in both directions (Demuth et al., 2016; Mikami et al., 2020); in GEJE, about 20% of non-evacuees stayed because past earthquakes had brought no large tsunami (Cabinet Office of Japan, 2012).	The channel changes speed.	Not supported
Panic	Panic flight is rare in disasters; emergency crowds cooperate (Quarantelli, 1954; Drury et al., 2009).	The channel slows everyone.	Not supported
Preparedness	Preparation and seeing evacuation signage raised the likelihood of evacuating in GEJE (Schmöcker et al., 2023); unprepared residents more often judged evacuation unnecessary (Cabinet Office of Japan, 2012).	The channel changes only shelter choice.	Not supported

D Reading and Asking

Each profile is described in three wordings that state the same seven facts in different forms and orders: a narrative ("A resident of a coastal town in Japan is in their fifties. Near home, this person does not know where the designated evacuation site is; ..."), a key–value list ("Profile: age 50s; had their work affected by the 2024 Noto earthquake; ..."), and a second narrative in the third person plural. Each is followed by one question and the instruction to reply with a single integer in the stated range. We use Qwen2.5-7B-Instruct quantized to 8 bits with MLX (the same weights used for reading). The greedy baseline decodes at most eight tokens; for the ICC in Table 13, 100 profiles are additionally asked five times each at temperature 0.8; all 5,100 answers contained a parseable integer. The continuous baseline computes a conditional expected value from next-token probabilities: one digit for the 1–7 scale, and a leading digit followed by a second digit or an end to the number for the other scales. It normalizes over in-range continuations; the two-digit approximation excludes 100 on the 0–100 scale. None of the greedy answers in these profiles or cue tests is 100. Prompts, answers, and analysis code are in `analysis/experiments/`. Table 13 summarizes greedy answers, Table 14 compares expected answers with read scores across wordings and cues, Table 15 gives the matched Seaside comparison, and Figure 14 shows what greedy answers do to the simulation.

Table 13: Direct prompting of Qwen2.5-7B-Instruct on 400 profiles versus RepE scores for the same profiles. Distinct: most different values given under any one wording; Shift: range of the population mean across the three wordings, as a share of the answer scale; ρ_{para}: mean Spearman correlation between wordings (†only one pair is defined, because one wording gave every profile the same answer); Same: share of profiles answered identically under all three wordings; ICC: between-profile share of variance under repeated sampling (T=0.8, 5 draws, 100 profiles); R^2: variance explained by the profile's attributes.

	Distinct						R^2	
Question	Prompt	RepE	Shift	ρ_{para}	Same	ICC	Prompt	RepE
Minutes after the warning before leaving (0–60)	3	400	10%	0.69†	63%	0.21	0.00	0.84
How strongly they act on the warning (1–7)	6	400	2%	0.65	52%	0.75	0.74	0.84
Likelihood of following neighbors (0–100)	7	400	25%	0.28	0%	0.30	0.61	0.79

Table 14: Mean pairwise Spearman correlation of resident scores across three wordings of the same 400 profiles. Expected answers use the conditional mean of numeric continuations up to two digits; greedy answers use the most likely generated integer. The one constant greedy-delay wording makes only one of its three pairwise correlations defined. Read scores use the same warning direction for both warning questions. Cue ordering is the share of paired high-pole descriptions that score above the low pole among 200 profile–cue pairs.

Question	Greedy answer	Expected answer	Read score
Departure delay	0.69†	0.55	0.73
Warning reliance	0.65	0.80	0.73
Following others	0.28	0.38	0.61
High cue > low cue: warning	67.5%	95%	100%
High cue > low cue: following	99.5%	100%	100%

Table 15: Matched Seaside runs with the warning and herd channels. Each interface uses the same 400 profiles, three profile wordings, ten seeds, and post-setup random state; scores are standardized within each wording. Entries are mean mortality (%); range is the spread of the wording means (pp) with a 95% interval from resampling seeds. Expected delay is sign-reversed to represent warning reliance.

Score source	Wording 1	Wording 2	Wording 3	Range
Read warning and herd	12.85	11.34	11.77	1.51 [1.2, 1.8]
Expected warning and herd	15.55	11.56	14.56	3.99 [3.7, 4.3]
Expected delay and herd	9.74	14.54	14.80	5.06 [4.8, 5.4]
Greedy warning and herd	15.24	10.88	18.58	7.70 [7.5, 8.0]

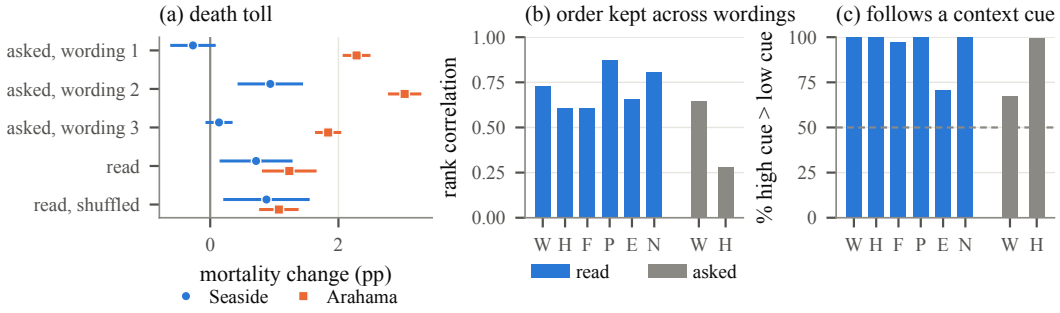

Figure 14: Diagnostic comparison with *greedy* direct answers. (a) Mortality when raw greedy answers or read scores for the same profiles drive warning and herd channels. These raw scales differ, so this panel is not a controlled comparison of methods. (b) Resident-rank correlations across three wordings (W, H, F, P, E, N: the six read factors). (c) High-pole cue ordering. The continuous expected-answer comparison is in Table 14.

E SECOND SCORE SET

The second score set covers the same 21,600 profiles and was produced by an earlier version of the reading pipeline, with a different probing protocol (Table 17). Several of its six columns coincide, so it carries fewer independent factors than ours. We run it through exactly the same simulations, seeds, and controls. Its results agree with ours on everything the shuffle test predicts: shuffling leaves the death toll unchanged while reassigning 41% of the victims in Seaside and 22% in Arahama. They differ in who is at risk, most visibly in the knowledge gradient discussed in §5 (Figure 15). Table 16 gives its channel effects in Seaside and Figure 15 its group differences; Table 17 gives the peak probe accuracies of the earlier protocol that produced it, which are not comparable with Figure 16.

Table 16: Seaside, second score set: paired differences from the matched baseline (mortality 13.16%, SD 0.70, over 10 seeds). ΔMort. in percentage points with 95% CI across seeds; ΔDep. in seconds (non-responders censored at 60 min); Non-resp.: share of residents who never leave; R^2: share of agent-level variance in departure time and in death (a 0/1 outcome, linear probability model) explained by the resident profile attributes.

Condition	Scores	ΔMort.	95% CI	ΔDep.	Non-resp. (%)	R^2_{dep}	R^2_{death}
Warning reliance	Read	+3.88	[+3.43, +4.34]	+256	12.0	0.69	0.36
	Linear	+3.66	[+3.23, +4.09]	+242	12.3	0.82	0.42
	Shuffled	+3.76	[+3.28, +4.23]	+262	12.2	0.01	0.01
Herd (+ warning)	Read	+1.51	[+1.03, +1.99]	+155	8.5	0.58	0.25
	Linear	+1.47	[+1.08, +1.85]	+141	8.8	0.72	0.31
	Shuffled	+1.43	[+0.95, +1.90]	+161	8.7	0.00	0.01
Family priority	Read	+0.50	[+0.32, +0.68]	+41	0.0	0.23	0.01
	Linear	+0.58	[+0.36, +0.79]	+41	0.0	0.29	0.01
	Shuffled	+0.27	[−0.05, +0.59]	+40	0.0	0.01	0.01
Preparedness	Read	−0.06	[−0.26, +0.13]	+0	0.0	0.01	0.01
	Linear	−0.15	[−0.32, +0.03]	+0	0.0	0.01	0.01
	Shuffled	−0.06	[−0.31, +0.19]	+0	0.0	0.01	0.01
Experience	Read	+1.54	[+1.20, +1.88]	+0	0.0	0.01	0.01
	Linear	+1.38	[+1.11, +1.66]	+0	0.0	0.01	0.01
	Shuffled	+1.57	[+1.34, +1.80]	+0	0.0	0.01	0.01
Panic	Read	+14.46	[+13.99, +14.92]	+0	0.0	0.01	0.03
	Linear	+14.18	[+13.87, +14.49]	+0	0.0	0.01	0.03
	Shuffled	+14.11	[+13.57, +14.66]	+0	0.0	0.01	0.01
All six	Read	+13.46	[+13.20, +13.73]	+196	8.5	0.56	0.15
	Linear	+13.03	[+12.62, +13.43]	+185	8.9	0.70	0.18
	Shuffled	+13.35	[+12.87, +13.83]	+201	8.7	0.01	0.01

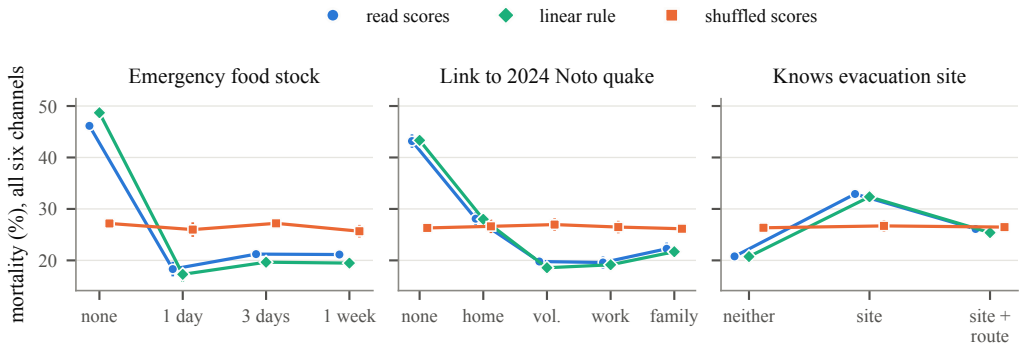

Figure 15: Second score set, Seaside: mortality by profile attribute with all six channels on (95% intervals over seeds). Compare Figure 4: the food-stock and earthquake-exposure gradients agree, the knowledge gradient is reversed.

Table 17: Peak pair-wise accuracy and the layer at which it occurs (counted from the output). n: held-out pairs. ✓: peak exceeds the layer-corrected one-sided threshold at $\alpha = 0.05$.

	Llama-2-7B				Qwen2.5-7B			
Factor	Peak	Layer	n	Readable	Peak	Layer	n	Readable
Experience	0.980	−19	256	✓	0.973	−5	256	✓
Herd influence	0.973	−15	256	✓	0.941	−9	256	✓
Preparedness	0.923	−13	65	✓	0.648	−1	256	✓
Warning reliance	0.703	−12	256	✓	0.638	−11	243	✓
Family priority	0.566	−29	256		0.641	−27	256	✓
Panic	0.590	−1	256		0.590	−5	256	

F Layer-wise Readability

Figure 16 shows held-out pair accuracy at every layer for our directions. The layer marked for each factor is chosen on validation wordings only (Appendix G); the shaded band lies below the one-sided threshold for chance at $\alpha = 0.05$, corrected for choosing among 28 layers.

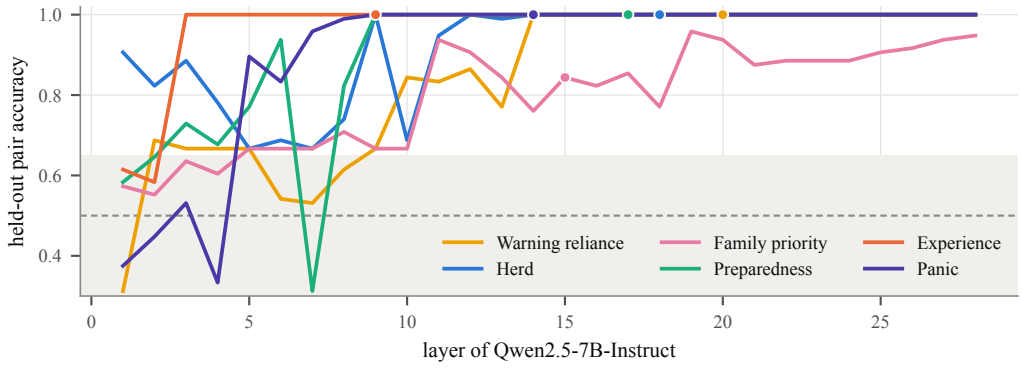

Figure 16: Held-out pair accuracy by layer, counted from the input (96 pairs per factor, no wording or resident shared with training). Dots: layer chosen on validation wordings.

G Scenario Construction

Each contrastive scenario combines a resident (24 personas that vary age, occupation, household, and location in a Japanese coastal town), a moment (4 descriptions of the earthquake and the warning), and one sentence expressing a pole of the factor (8 wordings per pole). The two scenarios of a pair share the resident and the moment and differ only in that sentence. Directions are fit on wordings 1–5 and 16 personas; the layer is chosen by the separation of wording 5 from a direction fit on wordings 1–4; accuracy is reported on wordings 6–8 and the remaining 8 personas (96 pairs per factor), so no held-out pair shares a wording or a resident with training. Every text is followed by the question "How will this person respond to the tsunami warning?" in the model's chat format, and hidden states are read at the last input token. Table 18 gives the first wording of each pole; all wordings are in `analysis/repe/scenarios.py`.

H Reproducibility

Simulations run in NetLogo 6.4 (Seaside) and 7.0.4 (Arahama) in headless mode from `analysis/experiments/run_grid.py`. Every run writes one record per resident with its profile attributes, scores, departure time, follower flag, outcome, score

Table 18: First wording of each pole, shown for one persona.

Factor	High pole	Low pole
Warning reliance	As soon as the official tsunami warning is announced, Aiko treats it as an order and starts leaving for high ground.	Aiko assumes the official warning is probably another false alarm and stays put for now.
Herd influence	Aiko looks out at the street and decides to leave only because the neighbors are already leaving.	Aiko decides on their own, regardless of what the neighbors are doing.
Family priority	Before evacuating, Aiko insists on first finding and gathering their family members.	Aiko evacuates directly, trusting that their family members will evacuate on their own.
Preparedness	Aiko grabs the emergency bag that is always packed and follows the evacuation route they have practiced.	Aiko has no emergency supplies ready and is not sure where the evacuation site is.
Experience	Aiko has evacuated from tsunami warnings several times before and knows how to move quickly.	Aiko has never experienced a tsunami warning and does not know what to expect.
Panic	Aiko panics, cannot think clearly, and keeps stopping in confusion.	Aiko stays calm, thinks clearly, and moves steadily.

source, seed, and channel configuration. `analyze.py` computes the original grid and population results; `gen_numbers.py` and `gen_tables.py` generate their manuscript values. `expected_answers.py` writes the numeric expectation data, `analyze_expected_answers.py` compares answer methods, and `fair_prompt_abm.py` runs and analyzes the standardized, paired Seaside comparison. Scripts for the traffic interventions and their result files are in the same directory.