

# Chiyoda: Entropy-Guided Information Control for Hazard-Coupled Pedestrian Evacuation

A Framework for Benchmarking Emergency Communication Efficacy

Gabriel Ong Zhe Mian

Independent

Singapore

gabrielzmong@gmail.com

## Abstract

Emergency evacuation is often treated as a physical routing problem in which warnings simply provide better information. The harder control problem is that the same message can improve local beliefs while synchronizing many agents toward a hazardous or capacity-constrained route. We present *Chiyoda*, an Information-Theoretic Evacuation Dynamics framework that treats emergency communication as a controllable safety action in a coupled physical-information system. Agents carry probabilistic exit and hazard beliefs, exchange distorted local gossip, receive beacon and responder messages, suffer physiological impairment under CBRN-like hazards, and route through belief-weighted paths under social-force crowd dynamics. Chiyoda evaluates static, global, responder-relay, entropy-targeted, density-aware, exposure-aware, and bottleneck-avoidance policies with telemetry that links belief entropy, belief accuracy, exposure, queueing, exit imbalance, and evacuation time. In a 50-seed primary study with 400 runs and 17780 intervention events, static\_beacon has the highest information-safety efficiency in the tested scenario (0.014), while global and adaptive policies show that broad reach and entropy reduction are not sufficient safety objectives. A 900-run robustness grid further shows that static local messaging remains the most information-safety-efficient tested policy across all hazard and familiarity regimes, even though the evacuation-count winner varies. A bounded LLM extension applies the same standard to generated guidance: validated sparse messages can be highly information-efficient, but equal-budget generated guidance weakens that advantage and does not eliminate harmful convergence. The central claim is therefore not that more information or richer language is always better, but that emergency communication must be evaluated by coupled belief and safety effects.

## Keywords

evacuation simulation, pedestrian dynamics, information theory, agent-based modeling, emergency communication, CBRN, entropy

Zenodo DOI: [10.5281/zenodo.19905070](https://doi.org/10.5281/zenodo.19905070)

## 1 Introduction

Emergency communication is usually treated as a benign input to evacuation: give people better information and their decisions should improve. That assumption is plausible at the individual level, but it is not obviously true at the crowd level. A warning that reduces uncertainty for everyone at once can also synchronize route choice, push agents toward the same exit, amplify a bottleneck, or

move people through a contaminated region before the physical hazard has dissipated. The question is therefore not simply whether information helps, but when, where, and for whom it helps.

This paper studies evacuation as a coupled physical-information system. The physical side includes layout geometry, social-force movement, bottleneck dynamics, hazard spread, and physiological impairment. The information side includes partial exit knowledge, hazard beliefs, signage and PA broadcasts, responder credibility, local gossip, message decay, and distortion. Chiyoda connects these layers by making agents plan through their believed world rather than the ground truth world. They may avoid a hazard they have heard about, miss an exit they do not know, herd when uncertainty is high, or follow a stale route because the belief has not yet decayed.

*Research question.* We ask when communication improves the belief-safety tradeoff and when it creates harmful convergence. More concretely: which intervention policies improve belief accuracy and reduce uncertainty per unit of induced exposure, queueing, and exit imbalance? The completed study package compares static signage, global broadcasts, responder relay, entropy-targeted broadcasts, density-aware targeting, exposure-aware targeting, and bottleneck-avoidance targeting, then stress-tests timing, budget, credibility, and repetition.

Our central hypothesis is that emergency communication should be evaluated as safety control, not as monotonic information delivery. Messages are therefore judged by their joint effect on belief quality and downstream movement risk: a policy can reduce uncertainty and still be unsafe if it concentrates agents at a bottleneck or routes them through hazard exposure. The same framing covers generated guidance: LLM-produced messages are evaluated only when they remain bounded, replayable, validator-checked, and scored with the same ISE and HCI diagnostics as deterministic messages.

*Contributions.* This paper makes five contributions.

- (1) **A controllable information-intervention layer.** Chiyoda turns emergency communication into a first-class policy object rather than a fixed scenario assumption. Policies can be static, authority-driven, or adaptive to entropy, density, exposure, and bottleneck pressure.
- (2) **A coupled ITED simulation loop.** Agents carry probabilistic beliefs over exits and hazards, update them through observation, beacon broadcast, responder relay, and gossip, then route through belief-weighted costs under social-force dynamics and hazard impairment.

- (3) **Information-safety metrics.** We introduce study outputs that jointly measure entropy reduction, belief accuracy gain, intervention reach, hazard exposure pressure, queue pressure, exit imbalance, information-safety efficiency, and harmful convergence.
- (4) **A reproducible empirical package.** YAML study definitions generate multi-seed bundles with Parquet/CSV tables and paper figures. The empirical evaluation uses a 50-seed primary study and two 30-seed supporting studies, plus a 900-run robustness grid across hazard severity and population familiarity. Generated statistics are ingested through `stats.tex`.
- (5) **A bounded generated-guidance extension.** Optional LLM-mediated guidance uses the same safety-control metrics, cache/replay requirements, and validation constraints. The extension tests whether generated messages improve downstream ISE/HCI, not whether they sound more natural.

Structure-wise, this paper first positions Chiyoda against pedestrian dynamics, information-aware evacuation, and ABM validation work. It then describes the model and intervention policies, explains the software pipeline, presents empirical results and robustness evidence, distinguishes the work from prior systems, states limitations, outlines future work, and concludes with the current evidence boundary.

## 2 Background

Pedestrian evacuation models have a long physical tradition. The social-force model represents pedestrians as self-driven particles subject to destination, repulsion, and contact forces [9]. Cellular automata models discretize space and update local movement decisions through transition rules [4]. Evacuation-model reviews emphasize that these physical models must be interpreted together with scenario assumptions, calibration data, and validation scope rather than as universally predictive engines [8, 23]. Fundamental-diagram work relates crowd density to walking speed and flow capacity [6, 30]. These tools are useful because evacuation failure is often physical: bottlenecks, counter-flow, density waves, and slower-is-faster effects can dominate individual intent.

Hazard-coupled evacuation adds another layer. In CBRN or smoke scenarios, the environment changes while people move through it; visibility, speed, and physiology can degrade as a function of exposure. Prior toxic-gas evacuation models already show that information perception and transmission matter for route choice under a spreading contaminant [15, 17]. Chiyoda builds on this insight but makes the communication policy itself the experimental object.

Information-aware evacuation models also intersect with route-choice theory, game theory, and bounded rationality. Pedestrian route-choice models connect activity constraints, perceived costs, and network structure [10], while evacuation experiments show that visibility can change route-selection behavior [7]. Bayesian formulations can model incomplete information and route-choice equilibria under congestion [16, 29]. Fire behavior and virtual-evacuation studies also show that exit choice is shaped by affiliation, social influence, stress, familiarity, and other occupants'

movement [2, 12, 14, 27]. Social group ABMs review the importance of communication, affiliation, and collective behavior, while also noting persistent gaps in behavioral realism and validation [25, 28]. Chiyoda takes a simulation-first position: it does not claim to solve empirical validation in one paper, but it makes belief uncertainty measurable and interventions reproducible.

The information-theoretic lens uses Shannon entropy [26]. For each agent, Chiyoda computes entropy over exit existence, hazard knowledge, and general danger. Global entropy is the population mean. Entropy is not a normative objective by itself: the goal is to test when reducing entropy improves safety and when it merely concentrates the crowd's behavior.

## 3 Model and Method

Chiyoda models evacuation as a feedback loop between physical state, information state, and action. At each step, hazards evolve; agents observe nearby exits and hazards; beacons, responders, gossip, and intervention policies update beliefs; agents revise intentions; belief-weighted navigation selects paths; social-force dynamics move agents; and telemetry records the macroscopic outcome.

### 3.1 Simulation State and Timing

The simulation state contains a grid layout, exits, hazards, responders, beacons, pedestrian agents, spatial indexes, and telemetry collectors. Each run is seeded and executes a fixed discrete-time horizon. Within a step, Chiyoda updates hazards before information propagation, then applies interventions before navigation. This ordering treats communication as a control action that can change the next route choice rather than as an after-the-fact label on the trajectory.

### 3.2 Agent Decision Loop

Agents are heterogeneous in exit familiarity, information confidence, susceptibility to gossip, and physiological response. An agent observes local state, receives messages, updates exit and hazard beliefs, chooses an intention, plans a route through its believed environment, and moves under a social-force update. The model therefore separates ground truth from perceived truth: two agents at the same location can choose different paths because they carry different exit knowledge, hazard beliefs, and message histories.

### 3.3 Belief State

Each agent carries a belief vector over exits, hazards, and general danger. Exit beliefs store existence probability, congestion estimate, freshness, source credibility, and gossip hop count. Hazard beliefs store position, severity estimate, radius estimate, freshness, credibility, and hop count. Agents do not route through the ground truth environment directly; they route through the subset of exits and hazards represented in their beliefs.

### 3.4 Route Choice and Hazard Coupling

Route choice is computed from the agent's current belief state. An agent first selects among exits whose existence probability exceeds the known-exit threshold, preferring exits with high existence probability, low congestion estimate, and fresh information. It then plans with  $A^*$  over the walkable grid. Each edge has a base traversal cost, a

density penalty from the current spatial index, and a hazard penalty from the agent’s hazard beliefs. Ground-truth hazards still determine exposure, visibility loss, physiological impairment, and direct observation, but they do not enter route planning unless the agent has observed or received a corresponding hazard belief. Agents with no known exit either explore or follow nearby movement depending on rationality and behavioral state. This separation is the main mechanism that lets us measure information as a causal intervention rather than as an oracle view of the environment.

### 3.5 Communication Channels

The baseline ITED channels are direct observation, beacon broadcast, and agent-to-agent gossip. Observation is range-limited by visibility and physiological impairment. Beacons provide high-credibility exit knowledge within a fixed radius. Gossip is local, probabilistic, and distorted by hop count and panic state.

Messages are not treated as perfect state synchronization. Each message carries source credibility, spatial reach, freshness, and content type. Recipients combine the message with their existing beliefs, which means communication can increase confidence, correct stale hazard estimates, or reinforce a route that later becomes congested.

### 3.6 Intervention Policies

The intervention policy interface is the experimental control surface. Every policy decides whether to fire, selects one or more spatial targets, builds a message, applies it to recipients inside a radius, and emits telemetry. Chiyoda currently implements seven deterministic baseline policies:

- **Static beacon:** periodic local messages from existing signage or PA beacon locations.
- **Global broadcast:** station-wide high-reach communication.
- **Responder relay:** mobile high-credibility messages emitted by released responders.
- **Entropy-targeted:** targets agents with highest belief entropy.
- **Density-aware:** targets dense local clusters.
- **Exposure-aware:** targets agents under high current or cumulative hazard pressure.
- **Bottleneck-avoidance:** targets active bottleneck zones.

The repository also includes an optional generated-message policy for bounded LLM experiments. This policy uses the same intervention hook as the deterministic policies, but separates target selection from message proposal: the simulator exposes a bounded state summary, a generator proposes structured route guidance, and a validator must accept the proposal before it can affect agent beliefs. This keeps the scientific comparison aligned with the safety-control hypothesis. A generated message is not treated as helpful because it is fluent; it is treated as a safety-control action whose downstream exposure, queueing, belief, and convergence effects must be measured.

Figure 1 summarizes the policy space. Figure 2 shows where the intervention hook sits in the simulation loop.

## 3.7 Information-Safety Metrics

The core coupled metric is information-safety efficiency:

$$\eta_{safe} = \frac{\Delta H^+ + \Delta A^+}{1 + P_{exposure} + P_{queue}},$$

where  $\Delta H^+$  is the sum of positive entropy reductions over intervention events,  $\Delta A^+$  is the sum of positive belief-accuracy gains,  $P_{exposure}$  is recipient-weighted hazard-load pressure, and  $P_{queue}$  is the sum of bottleneck queue pressure observed at intervention times. Runs with no interventions have ISE zero. ISE is deliberately not a utility function for all evacuation outcomes. It is a diagnostic for whether a communication policy purchases useful belief change cheaply in safety terms.

The complementary diagnostic is harmful convergence. HCI rises when entropy reduction coincides with concentrated exit usage, queueing, and exposure. A policy can therefore score well on ordinary information outcomes while scoring poorly on HCI if it makes agents more certain about the same constrained route. This is the main measurement distinction in the paper: Chiyoda evaluates communication by its coupled physical-information consequences, not by reach or entropy reduction alone.

In the current implementation, harmful convergence is computed as

$$\chi_{harm} = \frac{B_{exit}(1 + Q_{peak})(1 + \bar{E})}{1 + \Delta H_{int}^+},$$

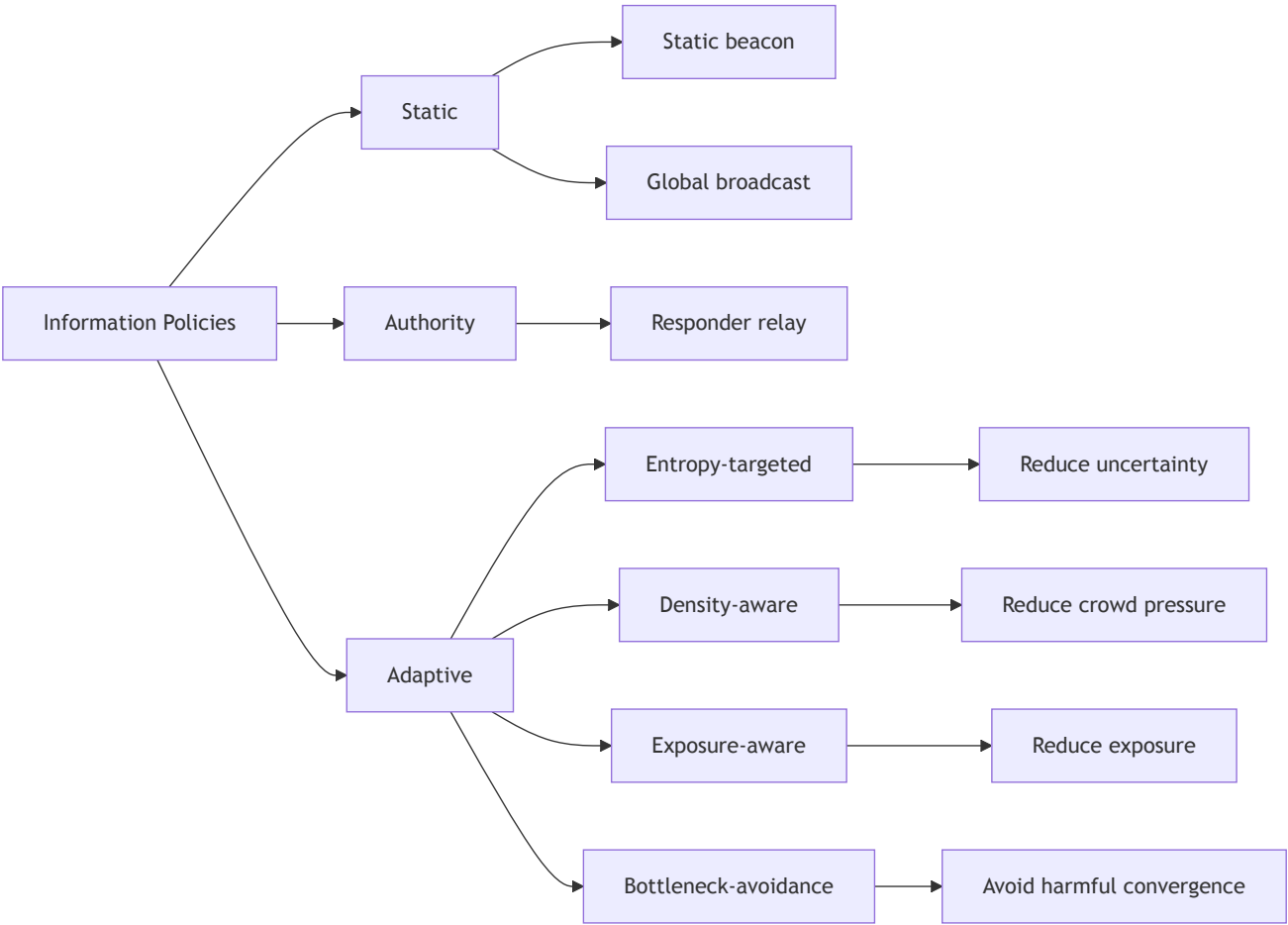
where  $B_{exit}$  is the final dominant-exit share,  $Q_{peak}$  is the peak bottleneck queue length,  $\bar{E}$  is mean cumulative hazard exposure, and  $\Delta H_{int}^+$  is the positive entropy reduction attributable to interventions. The denominator prevents ordinary information gain from being mistaken for harm, while the numerator raises the score when a population converges on a dominant exit under queueing and exposure pressure. Like ISE, HCI is a comparative study metric rather than a universal safety law.

## 4 Implementation

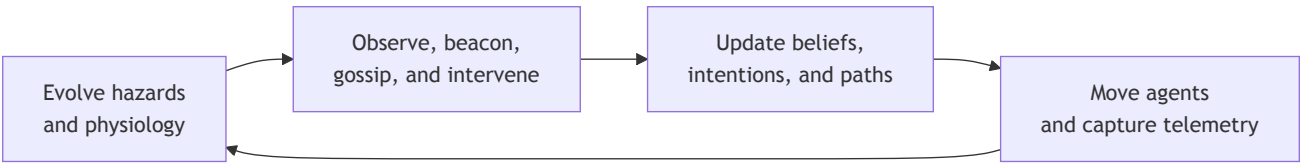
Chiyoda is implemented as a Python research toolkit. Scenario YAML files define layouts, cohorts, hazards, responders, information parameters, intervention policies, seeds, and variants. The scenario manager builds a simulation by loading the layout, constructing agents, hazards, exits, responders, navigation, spatial indexing, behavior state machines, and optional intervention policies. Cohort-level calibration fields expose exact walking speed, base rationality, credibility, gossip radius, and vision radius, while the behavior block exposes the full panic, recovery, freezing, and helping state-transition parameters used by the simulator.

The core simulation module records step telemetry, agent telemetry, cell-level maps, exit flow, bottleneck dynamics, hazard state, gossip events, and intervention events. The study runner materializes variants, runs each seed, aggregates tables, and exports study bundles. Analysis and figure export live in the analysis package.

The information intervention layer is deliberately small: policies produce target points and messages, while the shared executor applies messages to recipients and computes before/after entropy and accuracy. This keeps policy design separate from belief mutation and telemetry.



**Figure 1: Information intervention policy taxonomy. Policies differ by source, target signal, spatial reach, and safety objective.**



**Figure 2: Intervention hook in the ITED simulation loop. Policies read the current physical-information state, update recipient beliefs, and emit before/after telemetry before navigation.**

This organization is part of the research contribution rather than incidental software structure. Chiyoda is not only a crowd-motion simulator with extra logging; it exposes emergency communication as a controllable action with explicit targets, budgets, credibility, validation, and downstream safety telemetry. That separation enables the main empirical comparisons: message quality, recipient choice, route capacity, hazard exposure, and belief change can be varied independently instead of being collapsed into a single “informed agent” assumption.

The generated-message layer is optional and provider-neutral rather than a dependency of the baseline studies. It defines structured message requests, generated evacuation messages, validation records, a content-addressed JSON cache, deterministic template generation, replay-only generation, and a live OpenAI Responses API adapter. The `llm_guidance` policy uses cache-first semantics: if a matching cache record exists, the simulator reuses the cached message and validation result; otherwise a live provider may populate the cache in explicitly configured pilot runs. Paper-facing LLM runs should use replay-only mode after cache population.



The generated-message validator rejects proposals that invent exits, invent hazards, recommend already congested exits, give conflicting recommend/avoid instructions, exceed radius or credibility bounds, abstain, omit route guidance, produce vague text, or report confidence below the configured threshold. Intervention telemetry records the provider, model, cache key, cache status, validation status, validation reasons, and generated text so LLM behavior can be audited separately from downstream evacuation outcomes.

Two implementation details matter for reproducibility. First, A\* pathfinding uses per-step caches for density penalties, edge weights, and identical start/goal/belief queries. These caches are cleared whenever the simulation advances to a new navigation phase, so they do not change the cost function; they only avoid recomputing identical values during the same step. Second, gossip reports about the same physical hazard are merged into an existing hazard belief rather than appended as duplicate rumors. This is both a performance requirement and a modeling requirement: repeated reports should update credibility, freshness, severity, and radius estimates for the same hazard, not create an unbounded list of separate hazards.

The paper pipeline mirrors the code pipeline. Study execution emits tables and figures, `gen_stats.py` converts the exported bundle into `stats.tex`, and the LaTeX build injects those values into the abstract and evaluation. This keeps the manuscript tied to regenerated study artifacts rather than hand-maintained result counts.

For trajectory validation, Chiyoda exports per-agent step tables and includes a lightweight reference-comparison utility. The utility computes first-order summary differences such as agent count, duration, path length, displacement, speed, and local density from a Chiyoda bundle and a reference trajectory table. It is intentionally not a replacement for full trajectory-analysis libraries such as PedPy; its purpose is to make quick calibration checks repeatable inside the paper pipeline before deeper external analysis. The repository also includes a public Wuppertal bottleneck trajectory from the Pedestrian Dynamics Data Archive [3]. A dedicated validation script reads the PeTrack trajectory format, computes crossing times at the bottleneck measurement line used in the JuPedSim/PedPy example, and compares crossing count, mean flow, and time headway against a Chiyoda bottleneck proxy scenario. This adds an external measurement target while keeping the claim narrow: it validates the comparison pipeline and exposes calibration gaps, not operational station prediction. A small follow-up sweep over bottleneck width, exit width, base speed, density slowdown, and social-force repulsion records the current best proxy but keeps the result outside the set of calibrated behavioral claims.

## 5 Evaluation

We evaluate Chiyoda with one primary policy comparison and two supporting stress studies. The primary study is a 50-seed comparison of eight emergency communication regimes: no deliberate intervention, static beacon broadcast, global PA broadcast, responder relay, entropy-targeted communication, density-aware communication, exposure-aware communication, and bottleneck-avoidance communication. The exported bundle contains 400 runs, 8 variant aggregates, and 17780 intervention events. Each run uses the same station-sarin scenario, a 120-agent commuter/tourist population,

asymmetric initial information, gossip, beacon infrastructure, three delayed responders, and a 500-step horizon.

The two supporting studies each use 30 seeds. The intervention ablation study varies targeting budget, timing, and bottleneck awareness. The message-quality study varies broadcast credibility, delay, and repetition. These studies do not replace the main comparison; they test whether the main information-safety interpretation survives natural design changes.

### 5.1 Outcome Measures

We report three classes of outcome. Physical outcomes include evacuated agents, mean travel time, incapacitation count, mean hazard exposure, peak bottleneck queue, and exit imbalance. Information outcomes include mean entropy, entropy reduction, intervention reach, entropy reduction attributable to interventions, and belief-accuracy gain. Coupled outcomes include information-safety efficiency (ISE) and harmful convergence index (HCI). These coupled metrics are the main object of the paper: they ask whether a policy buys useful belief improvement without creating exposure or queueing pressure.

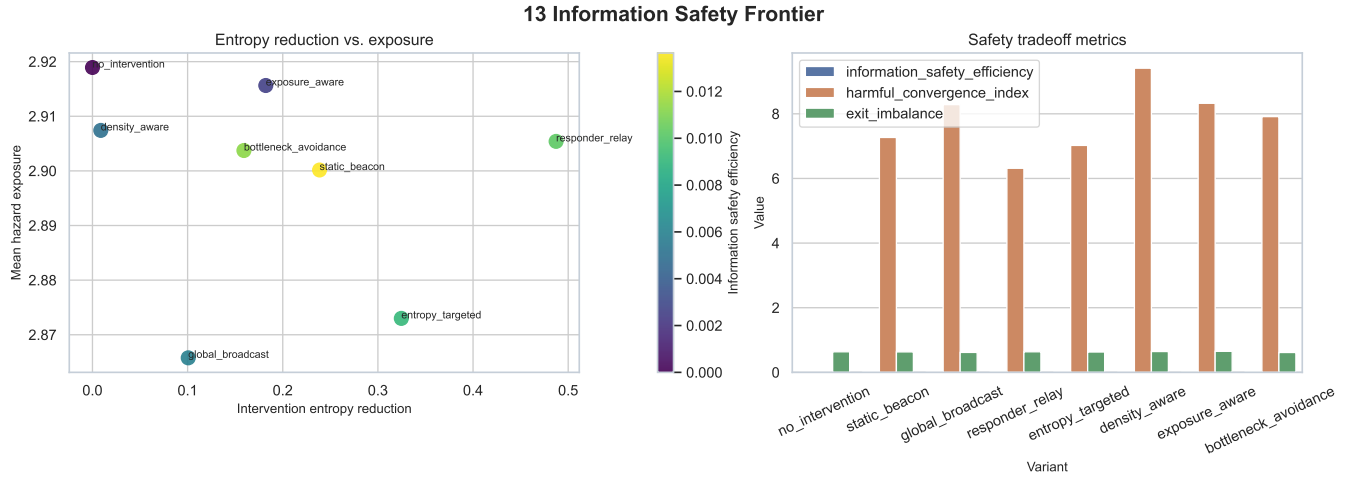
Variant tables report means across seeds. Pairwise comparisons against the no-intervention baseline use two-sided Mann-Whitney tests over run-level summary rows. We treat  $p < 0.05$  as evidence of a seed-level distributional difference in this simulator, and we use effect sizes as descriptive support rather than as standalone proof of operational superiority. This choice is conservative for the current stage: the simulator is stochastic and stylized, so the evaluation focuses on tradeoff patterns that recur across studies rather than on declaring a universal winner.

### 5.2 Primary Policy Comparison

Table 1 summarizes the 50-seed variant aggregates. The first result is negative but important: no communication policy dominates the baseline across physical evacuation outcomes. The no-intervention baseline evacuates 10.76 agents on average. Static beacons evacuate 10.90, global broadcast evacuates 11.00, and the remaining active policies range from 9.70 to 10.86. Mann-Whitney tests against the baseline do not show significant improvements in evacuated count or hazard exposure at  $p < 0.05$  for any active policy at  $n = 50$  seeds. These results rule out the stronger claim that more information monotonically improves evacuation.

The second result is that communication policy still matters, even when physical outcomes remain noisy. Static beacons have the highest information-safety efficiency in the main scenario (static\_beacon, 0.014). They combine local reach, repeated exposure, and high credibility: enough to reduce belief uncertainty without synchronizing the full population toward the same capacity-constrained route. Responder relay has lower efficiency than static beacons, but also the lowest HCI among the active policies in the main study (6.31), which supports the idea that delayed, mobile, high-credibility sources can be conservative information controllers in this setting.

The third result is that wide reach is not the same as useful reach. Global broadcast slightly improves mean evacuation count relative to the baseline (11.00 versus 10.76) and has the lowest mean hazard exposure among the main policies (2.866). However, it ranks fifth by ISE (0.0058) and has a high HCI (8.28). In this scenario,



**Figure 3: Information-safety frontier across communication policies in the 50-seed main study. Static beacons produce the highest information-safety efficiency, while density-aware and exposure-aware policies show the largest harmful-convergence penalties under this scenario.**

**Table 1: Variant aggregates from the 50-seed information-control study. ISE denotes information-safety efficiency; HCI denotes harmful-convergence index.**

| Policy               | Evac. | Travel (s) | Exposure | Queue | ISE    | HCI  |
|----------------------|-------|------------|----------|-------|--------|------|
| Static beacon        | 10.90 | 10.88      | 2.900    | 2.72  | 0.0136 | 7.27 |
| Bottleneck avoidance | 9.70  | 9.76       | 2.904    | 2.84  | 0.0113 | 7.91 |
| Responder relay      | 10.52 | 10.69      | 2.905    | 2.76  | 0.0102 | 6.31 |
| Entropy-targeted     | 9.90  | 10.31      | 2.873    | 2.86  | 0.0091 | 7.02 |
| Global broadcast     | 11.00 | 11.86      | 2.866    | 2.84  | 0.0058 | 8.28 |
| Density-aware        | 10.78 | 10.63      | 2.907    | 2.80  | 0.0050 | 9.41 |
| Exposure-aware       | 10.86 | 11.65      | 2.916    | 2.86  | 0.0026 | 8.33 |
| No intervention      | 10.76 | 11.53      | 2.919    | 2.74  | 0.0000 | 0.00 |

the global message is useful but inefficient: it homogenizes route guidance broadly, while local and capacity-aware policies produce more information benefit per unit of safety pressure.

The fourth result is that adaptive targeting exposes a real control tradeoff. Entropy targeting has the largest intervention accuracy gain in the main study (1.76), and both entropy targeting and bottleneck avoidance show seed-level travel-time differences against no intervention ( $p = 0.035$  and  $p = 0.014$ , respectively). Yet these same policies do not improve evacuation count, and bottleneck avoidance trends toward fewer evacuated agents ( $-1.06$ ,  $p = 0.061$ ). The practical interpretation is that better-informed trajectories can be shorter for agents who move, while still failing to improve aggregate completion under capacity and hazard pressure.

### 5.3 Intervention Ablation

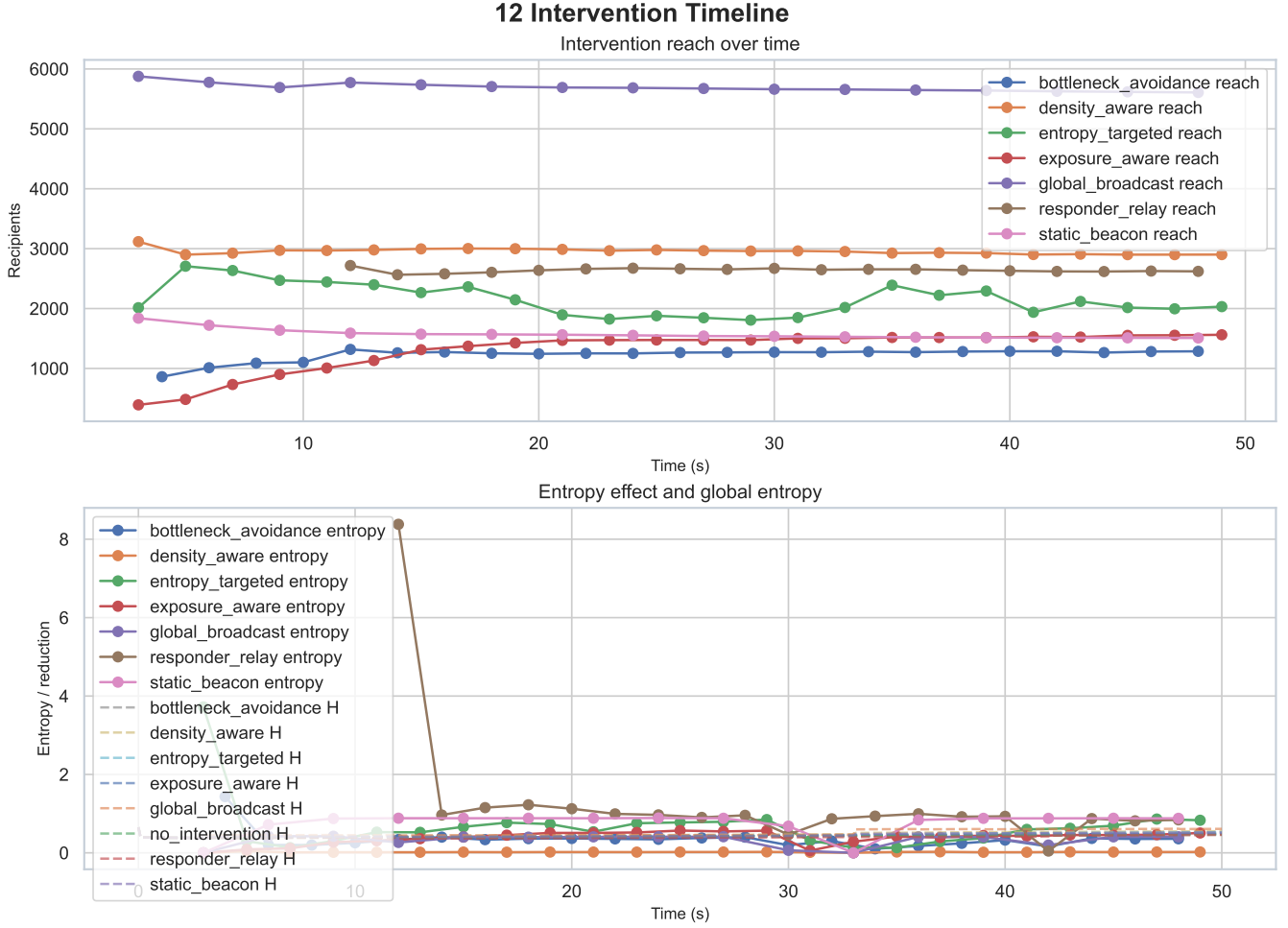
Table 2 reports the 30-seed ablation study. The strongest finding is budget non-monotonicity. Early low-budget entropy targeting has the highest ISE in the ablation (0.0253), more than double early high-budget entropy targeting (0.0092). The high-budget version improves belief accuracy more strongly (1.89 versus 1.26), but also

produces larger queue pressure and lower information-safety efficiency. This supports the mechanism behind our hypothesis: more certainty is not automatically safer when many agents receive similar guidance under shared bottleneck constraints.

The timing ablation also matters. Late entropy targeting recovers evacuation count relative to early entropy targeting (11.47 versus 10.70 under the high-budget setting), but its ISE is lower. Bottleneck targeting is less sensitive to late deployment: the late bottleneck condition has slightly higher evacuation count and slightly higher ISE than early bottleneck targeting. This suggests that bottleneck-aware control is a better candidate for delayed response settings in this simulator, while entropy targeting needs stricter budget control.

### 5.4 Message Quality

Table 3 reports the 30-seed message-quality study. Credibility has the expected direction for information efficiency: high-credibility global broadcast has the highest ISE among the broadcast-quality conditions (0.0066), delayed high-credibility broadcast is close behind (0.0063), medium credibility falls to 0.0034, and low credibility falls to 0.0026. Repeating low-credibility messages frequently does



**Figure 4: Intervention reach and entropy effects over time. Broad reach and useful safety effect separate: global broadcast reaches many agents, but its information-safety efficiency remains below local static beacons, bottleneck avoidance, responder relay, and entropy targeting.**

**Table 2: Supporting intervention ablation over 30 seeds.**

| Condition                  | Evac. | Travel (s) | Exposure | Queue | ISE    | HCI  |
|----------------------------|-------|------------|----------|-------|--------|------|
| Entropy early, low budget  | 10.30 | 11.74      | 2.809    | 2.93  | 0.0253 | 8.57 |
| Bottleneck late            | 11.17 | 11.57      | 2.873    | 2.77  | 0.0143 | 7.69 |
| Bottleneck early           | 10.40 | 10.85      | 2.778    | 2.97  | 0.0140 | 7.98 |
| Entropy early, high budget | 10.70 | 11.65      | 2.879    | 3.00  | 0.0092 | 7.08 |
| Entropy late, high budget  | 11.47 | 11.80      | 2.866    | 2.80  | 0.0070 | 7.43 |
| No intervention            | 11.20 | 11.84      | 2.869    | 2.67  | 0.0000 | 0.00 |

not repair this loss; it has the lowest ISE in the study (0.0009) while preserving high harmful convergence (8.92).

This support study strengthens the interpretation of global broadcast. Global messages can increase completed evacuations in some parameterizations, but the information-safety metric penalizes conditions where the same broad signal also creates route synchronization. Message quality therefore behaves like a control parameter, not

a cosmetic parameter: credibility and timing change whether communication improves useful beliefs or merely coordinates crowd movement.

## 5.5 Hypotheses Revisited

H1 predicted that adaptive targeting would improve information-safety efficiency over global broadcast. The 50-seed result supports

**Table 3: Supporting message-quality study over 30 seeds.**

| Condition                 | Evac. | Travel (s) | Exposure | Queue | ISE    | HCI  |
|---------------------------|-------|------------|----------|-------|--------|------|
| High-credibility global   | 11.13 | 12.01      | 2.889    | 2.83  | 0.0066 | 8.42 |
| Delayed high credibility  | 10.80 | 10.70      | 2.856    | 2.80  | 0.0063 | 8.09 |
| Medium-credibility global | 11.53 | 10.85      | 2.862    | 2.80  | 0.0034 | 7.12 |
| Low-credibility global    | 12.10 | 11.24      | 2.893    | 2.90  | 0.0026 | 9.07 |
| Frequent low credibility  | 11.93 | 11.79      | 2.858    | 2.80  | 0.0009 | 8.92 |
| No intervention           | 11.20 | 11.84      | 2.869    | 2.67  | 0.0000 | 0.00 |

this for static beacons, bottleneck avoidance, responder relay, and entropy targeting, all of which exceed global broadcast by ISE. The result is weaker for density-aware and exposure-aware targeting, which have nonzero efficiency but lower safety efficiency than global broadcast in this scenario.

H2 predicted that global broadcast would reduce entropy quickly but risk route synchronization. The data support the synchronization concern more strongly than the entropy-reduction claim. Global broadcast has high reach and high HCI, but its ISE trails more localized policies. The message-quality study further shows that frequent low-credibility broadcast is especially inefficient.

H3 predicted that exposure-aware targeting would reduce hazard exposure. Exposure-aware targeting does not meaningfully reduce exposure in the 50-seed main study: its mean hazard exposure is 2.916, close to the baseline value of 2.919 and above several other policies. This hypothesis is not supported by the present configuration.

H4 predicted that responder relay would be robust but delayed. The result is partially supported. Responder relay is third by ISE, has the lowest HCI among the main active policies, and avoids the extreme synchronization penalties of some adaptive policies. Its limitation is not obvious physical superiority, but conservative information control.

## 5.6 Interpretation

The results support our hypothesis that emergency communication is a coupled belief-and-safety control problem. Information interventions should be evaluated as safety-control actions, not as monotonic entropy minimizers. A policy can reach many agents and still have low safety efficiency; another can improve belief accuracy and still fail to improve aggregate completion. The operational implication is not simply to broadcast more, but to broadcast credibly, locally, and with awareness of capacity and exposure. Static beacon and responder relay policies currently look like conservative baselines, while entropy targeting is scientifically valuable because it exposes the failure mode created by overconfident, synchronized route choice.

## 5.7 Bounded LLM Guidance Extension

The LLM extension tests generated evacuation messages as constrained control actions, not as unconstrained advice. The simulator exposes a structured state summary, asks the provider for recommended exits, avoided exits, hazard references, confidence, and optional abstention, and validates the proposal before it can affect agent beliefs. Paper-facing runs use cache-first or replay-only

execution, record model metadata and validation outcomes, and compare generated guidance with deterministic policies by ISE and HCI rather than by fluency.

**5.7.1 Medium LLM Extension Result.** A first medium-scale LLM extension study applies this protocol to 80 runs: eight policy variants across ten random seeds. The study includes the deterministic baselines, a deterministic template generator, three live OpenAI prompt/validator ablations, and a replay-only verification variant. The live LLM policies make only eight intervention attempts per run, compared with 32–72 interventions for the deterministic communication policies.

The result is useful but bounded. The live OpenAI variants achieve higher information-safety efficiency than the deterministic policies under this intervention budget, and replay exactly matches the safety-strict live variant. However, the harmful-convergence index remains high. The extension therefore supports a cautious claim: validated generated messages can be made replayable and efficient, but richer adaptive language does not by itself solve crowd synchronization risk.

**5.7.2 Target-Selection Ablation.** The medium LLM result has an important confound: the generated-message policies use a much smaller recipient budget than the deterministic policies. To separate generated-message behavior from recipient selection, we ran a target-selection ablation with deterministic template generation. The provider, prompt style, validator profile, cadence, radius, and budget are held fixed; only the target selector changes.

The ablation shows that target selection is not a minor implementation detail. Bottleneck, entropy, density, and static target selectors all produce much higher ISE than the deterministic baselines under a small generated-message budget, while global and exposure targeting fall close to the deterministic frontier. Density targeting has the best evacuation count among the generated variants and the lowest generated-policy HCI, but it still does not beat the static deterministic beacon on harmful convergence. The implication is that the LLM extension should not be interpreted as a language-only improvement. The safety effect is created by a coupled design: bounded messages, validation, small recipient budgets, and target selection.

**5.7.3 Regime Robustness.** We therefore ran a focused LLM robustness extension over the same hazard-severity and population-familiarity axes used in the deterministic regime grid. The extension contains nine regimes, five seeds per regime, one live OpenAI safety-strict cache-population variant, and one replay-only variant per



| Policy                 | Evacuated | ISE    | HCI  | Recipients |
|------------------------|-----------|--------|------|------------|
| Static beacon          | 11.0      | 0.0145 | 6.81 | 503.7      |
| Entropy targeted       | 9.2       | 0.0079 | 8.10 | 1064.3     |
| Bottleneck avoidance   | 9.4       | 0.0124 | 7.55 | 585.7      |
| Template safety        | 9.0       | 0.0155 | 7.86 | 446.6      |
| OpenAI state-only      | 9.0       | 0.0414 | 8.39 | 132.3      |
| OpenAI safety-strict   | 9.9       | 0.0491 | 7.84 | 133.8      |
| OpenAI entropy-lenient | 10.0      | 0.0429 | 8.76 | 138.8      |
| Replay safety-strict   | 9.9       | 0.0491 | 7.84 | 133.8      |

**Table 4: Medium LLM extension summary over ten seeds. LLM guidance improves information-safety efficiency mainly by operating under a much smaller intervention budget, but it does not yet reduce harmful convergence relative to the strongest conservative baseline.**

| Target selector | Evacuated | ISE    | HCI  | Recipients |
|-----------------|-----------|--------|------|------------|
| Bottleneck      | 9.2       | 0.0450 | 8.48 | 85.4       |
| Entropy         | 9.7       | 0.0422 | 8.66 | 128.5      |
| Density         | 10.2      | 0.0391 | 7.35 | 146.3      |
| Static          | 10.0      | 0.0354 | 7.84 | 95.5       |
| Global          | 9.3       | 0.0106 | 8.58 | 67.7       |
| Exposure        | 9.3       | 0.0087 | 8.45 | 72.1       |

**Table 5: LLM target-selection ablation over ten seeds using deterministic template generation. Message-generation settings are fixed; only recipient selection changes.**

regime. Each run uses the same eight-attempt intervention budget as the medium safety-strict LLM policy.

The live run produced 360 generated-message events. Safety validation accepted 359 OpenAI messages and rejected one congested-exit recommendation, which was handled by the deterministic fallback. Replay reproduced the live aggregate outcomes exactly across evacuated count, ISE, HCI, intervention count, and recipient count. The robustness check strengthens the reproducibility claim for generated guidance, but it also sharpens the limitation: HCI rises sharply with hazard severity. LLM guidance remains safety-efficient under a small budget, but it still does not solve harmful convergence in severe hazard regimes.

The direct comparison against deterministic robustness policies clarifies the tradeoff. Under the smaller LLM intervention budget, safety-strict generated guidance has higher ISE than every deterministic communication policy averaged across the regime grid. It does not, however, lower HCI relative to the static beacon, bottleneck-avoidance, or entropy-targeted baselines. This is consistent with the target-selection ablation: bounded generated guidance can buy useful belief change efficiently, while the remaining safety problem is still route convergence under hazard pressure.

**5.7.4 Objective and Budget Ablations.** The expanded LLM evidence tests whether this efficiency result is a language effect, a prompt-objective effect, or primarily a budget effect. The prompt-objective ablation keeps target selection, validator, cadence, radius, budget, provider, and model fixed while varying the live OpenAI prompt between safety, hazard avoidance, anti-convergence, and urgency framings.

Prompt objectives measurably change the downstream tradeoff, but not in the simple way a language-only interpretation would predict. The safety prompt has the highest generated-policy ISE. The

hazard-avoidance prompt has lower HCI than the safety prompt, but also slightly lower ISE. The anti-convergence prompt does not reduce HCI in this setting. All 320 live OpenAI messages in this ablation passed validation and replay reproduced the live aggregates, which makes the result a controlled prompt-objective comparison rather than a cache artifact.

The budget-equivalence sweep is the sharper test. Sparse OpenAI guidance keeps high ISE under only eight interventions per run. When generated guidance is given static-beacon or entropy-targeted intervention budgets, its ISE falls substantially. The equal-budget variants reduce HCI relative to sparse LLM guidance, but they do not dominate the conservative deterministic baselines. This supports the bounded LLM claim: validated generated guidance can be a high-efficiency sparse intervention, but it is not a drop-in replacement for deterministic safety-control policies.

Seed-level tests point in the same direction. Sparse safety-prompt guidance has higher ISE than the static beacon, entropy-targeted, and bottleneck-avoidance baselines in the prompt-objective study under the reported Mann-Whitney comparisons. The same comparisons do not show an HCI improvement. The budget-equivalence comparisons show the reverse pressure: larger generated-message budgets lower HCI descriptively, but weaken the sparse LLM ISE advantage. These results characterize LLM guidance as a bounded generative extension: useful because it evaluates generated emergency messages as safety-control actions, but still limited by route-convergence risk and intervention-budget sensitivity.

## 5.8 What the Evidence Supports

The empirical package supports five claims. First, emergency communication behaves as a safety-control action: changing who receives a message, when they receive it, and how credible it is

| Hazard | Evacuated | ISE    | HCI   |
|--------|-----------|--------|-------|
| Low    | 10.00     | 0.0417 | 3.75  |
| Medium | 9.27      | 0.0434 | 8.93  |
| High   | 9.33      | 0.0293 | 18.92 |

**Table 6: Focused LLM regime robustness extension, averaged across low, mixed, and high population-familiarity regimes. The replay-only variants match the live OpenAI aggregate metrics exactly.**

| Deterministic policy | Det. ISE | LLM/Det. ISE | Det. HCI | LLM HCI $\Delta$ |
|----------------------|----------|--------------|----------|------------------|
| Static beacon        | 0.0128   | 2.97         | 9.41     | +1.12            |
| Bottleneck avoidance | 0.0100   | 3.82         | 9.98     | +0.55            |
| Entropy targeted     | 0.0089   | 4.26         | 8.38     | +2.15            |
| Global broadcast     | 0.0063   | 6.09         | 10.60    | -0.07            |

**Table 7: Direct comparison between deterministic robustness policies and the live OpenAI safety-strict LLM extension, averaged across all nine hazard-familiarity regimes. The deterministic grid uses 20 seeds per cell; the LLM extension uses five seeds per cell. Positive HCI deltas mean the LLM extension has higher harmful convergence.**

| Policy                  | Evacuated | ISE    | HCI  | Recipients |
|-------------------------|-----------|--------|------|------------|
| Static beacon           | 11.0      | 0.0145 | 6.81 | 503.7      |
| Entropy targeted        | 9.2       | 0.0079 | 8.10 | 1064.3     |
| Bottleneck avoidance    | 9.4       | 0.0124 | 7.55 | 585.7      |
| OpenAI safety           | 9.1       | 0.0444 | 8.37 | 130.9      |
| OpenAI hazard avoidance | 9.6       | 0.0426 | 7.83 | 133.1      |
| OpenAI anti-convergence | 9.3       | 0.0409 | 8.72 | 129.6      |
| OpenAI urgency          | 9.7       | 0.0398 | 8.30 | 124.2      |

**Table 8: Live OpenAI prompt-objective ablation. Target selection, validator, cadence, radius, budget, provider, and model are held fixed for the generated guidance variants.**

| Policy                    | Evacuated | ISE    | HCI  | Recipients |
|---------------------------|-----------|--------|------|------------|
| Static beacon             | 12.4      | 0.0177 | 6.31 | 492.6      |
| Entropy targeted          | 9.4       | 0.0078 | 8.62 | 1016.8     |
| Bottleneck avoidance      | 10.0      | 0.0145 | 7.48 | 551.0      |
| OpenAI sparse             | 11.2      | 0.0449 | 8.82 | 122.0      |
| OpenAI static-equivalent  | 10.2      | 0.0116 | 8.16 | 732.0      |
| OpenAI entropy-equivalent | 10.4      | 0.0086 | 7.78 | 978.4      |

**Table 9: Live OpenAI budget-equivalence sweep. Sparse generated guidance is compared with generated guidance using static-beacon and entropy-targeted intervention budgets.**

changes exposure, queueing, exit concentration, and belief state jointly. Second, local static messaging is the highest-ISE tested deterministic baseline in both the 50-seed main comparison and the 900-run robustness grid. Third, broad reach is not equivalent to safety-efficient reach; global broadcast can improve completed evacuations in some regimes while still delivering less belief improvement per unit of exposure and queue pressure. Fourth, entropy and belief-accuracy gains are not sufficient objectives, because they can coordinate many agents toward the same constrained route. Fifth, validated sparse generated guidance can buy belief improvement efficiently, but its advantage depends on recipient budget and does not remove harmful-convergence risk.

The same evidence rules out stronger claims. The experiments do not show that static beacons are a universal operational optimum,

because the evacuation-count winner changes across hazard and familiarity regimes. They also do not establish operational station predictive validity. The repository now includes a public Wuppertal bottleneck trajectory check that compares crossing counts, mean flow, and time headways against a Chiyoda bottleneck proxy. The best sweep candidate uses a seven-cell proxy bottleneck, 1.60 m/s base speed, reduced density slowdown, and baseline social-force repulsion; it still reaches only 49 of 75 reference crossings, with mean flow 0.675 versus 1.163 ped/s and mean headway 1.513 versus 0.871 s. That check is therefore a calibration diagnostic rather than validation of station layouts, hazard physics, or incident response. The LLM studies also do not show that generative or personalized

| Claim                             | ISE $\Delta$ | ISE $p$ | HCI $\Delta$ | HCI $p$ |
|-----------------------------------|--------------|---------|--------------|---------|
| Sparse safety vs static           | +0.0299      | 0.0002  | +1.55        | 0.3447  |
| Sparse safety vs entropy          | +0.0365      | 0.0002  | +0.27        | 0.9097  |
| Sparse safety vs bottleneck       | +0.0320      | 0.0002  | +0.82        | 0.6232  |
| Hazard prompt vs safety prompt    | -0.0018      | 0.7913  | -0.54        | 0.8501  |
| Anti-convergence vs safety prompt | -0.0035      | 0.4274  | +0.35        | 0.8205  |
| Static budget vs sparse LLM       | -0.0332      | 0.0079  | -0.67        | 1.0000  |
| Entropy budget vs sparse LLM      | -0.0363      | 0.0079  | -1.05        | 0.4206  |

**Table 10: Seed-level Mann–Whitney comparisons for the main LLM extension claims. Positive ISE deltas favor the test variant; negative HCI deltas favor the test variant.**

messages are safe by default, because they find efficiency gains without a corresponding HCI win over the conservative deterministic baselines.

### 5.9 Threats to Evaluation

The current evidence is stronger than the pilot run, but still not final external validation. The experiments use one station geometry, one hazard scenario, one population size, and a fixed evacuation horizon. The supporting studies vary intervention design but not architectural layout, sensor latency, or calibrated pedestrian demographics. The HCI and ISE metrics are also model constructs; future work should test alternative normalizations and compare them against trajectory-level bottleneck events, drill data, or expert-coded evacuation plans.

## 6 Regime Robustness Study

The primary evaluation uses one station layout, one CBRN-like hazard placement, one mixed-familiarity population, and a fixed evacuation horizon. That is enough to identify the information-safety tradeoff, but not enough to claim that the observed policy ordering generalizes. The regime robustness study therefore treats external validity as the object of study.

The robustness study asks where each communication policy becomes more efficient, less efficient, or more convergence-prone under the model. It varies three hazard regimes, three population familiarity regimes, and five representative communication policies over 20 seeds, yielding 900 runs. The policy set is intentionally smaller than the main study: no intervention, static beacons, global broadcast, entropy-targeted communication, and bottleneck-avoidance communication. These policies span the main theoretical contrast between no control, local fixed control, global high-reach control, uncertainty-driven adaptive control, and capacity-aware adaptive control.

This study is not designed to maximize a single score. Its role is to estimate an operating envelope: when local communication remains efficient, when global broadcast becomes wasteful or risky, when entropy targeting needs budget control, and when bottleneck-aware messaging is most valuable.

### 6.1 Regime Results

Table 12 summarizes the completed robustness grid. The clearest result is that static beacons have the highest information-safety efficiency in all nine hazard–familiarity regimes. This does not mean static beacons maximize every evacuation endpoint. It means

that local, repeated, high-credibility messaging consistently buys belief improvement with less exposure and queue pressure than broader or more adaptive alternatives in this simulator family.

Averaged across the nine regimes, static beacons have the highest ISE (0.0128), followed by bottleneck avoidance (0.0100), entropy targeting (0.0090), and global broadcast (0.0063). The global/static ISE ratio ranges from 0.27 to 0.69, so broad reach remains a weak proxy for safety-efficient communication even when global broadcast sometimes produces the largest evacuation count. This is the key robustness result: the information-safety ranking is more stable than the raw evacuation-count ranking.

The robustness grid also clarifies the adaptive policies. Entropy targeting has the largest belief-accuracy gain in every regime, but it never has the best ISE. Its value is therefore diagnostic: it exposes the gap between belief improvement and safety-efficient control. Bottleneck avoidance is usually the second-best active policy by ISE, especially in low-familiarity regimes, but it does not dominate static beacons. Under high hazard pressure, all active policies exhibit much larger harmful-convergence values than under low hazard pressure, which confirms that communication alone cannot erase the physical constraints imposed by the hazard field and bottlenecks.

### 6.2 Interpretation Protocol

The robustness grid is interpreted as a claim filter rather than as a leaderboard. A policy is treated as *robust within the tested grid* only if its information-safety efficiency remains positive and competitive across most hazard–familiarity cells without producing consistently high harmful convergence. A policy is treated as *regime dependent* if it performs well only under particular hazard severity or familiarity conditions. A policy is treated as *inefficient under this model* if it improves belief state while repeatedly increasing exposure, queue pressure, or exit concentration.

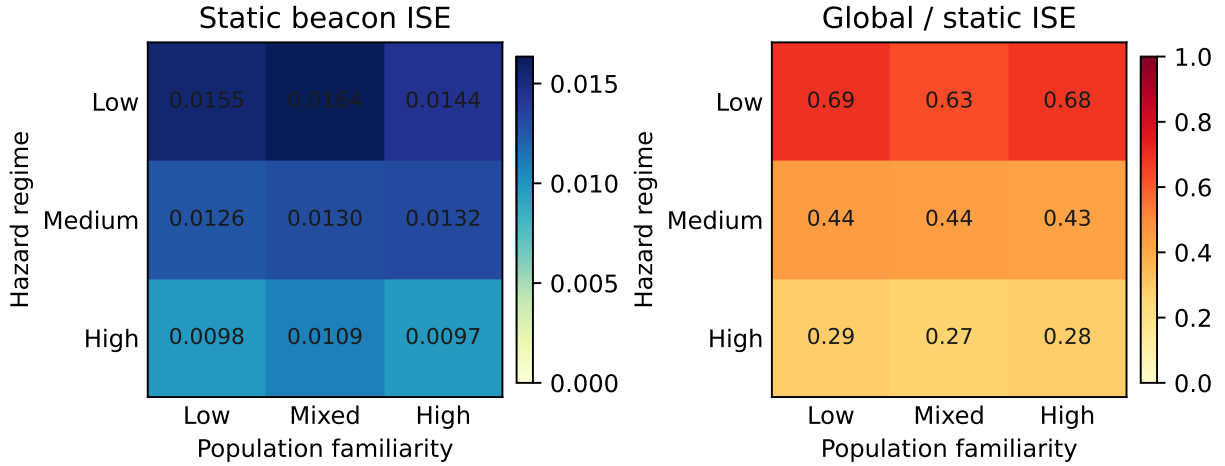
This protocol matters because the contribution is a measurement framework, not a universal evacuation prescription. Static beacons are the highest-ISE conservative controller across the tested regimes, while global broadcast, entropy targeting, and no intervention each win some cells by completed evacuations. Similarly, entropy targeting is scientifically important even though it is not operationally dominant, because it reveals when uncertainty reduction turns into synchronized route choice. We therefore report both policy rankings and regime-specific failure modes.

**Table 11: Regime robustness matrix. The full factorial design contains  $3 \times 3 \times 5 \times 20 = 900$  seeded runs.**

| Axis                   | Levels                                    | Purpose  |
|------------------------|---|--|
| Hazard regime          | Low, medium, high severity/spread         | Tests whether information control changes under stronger physical threat.                    |
| Population familiarity | Low, mixed, high exit familiarity         | Tests whether communication helps unfamiliar crowds and whether familiar crowds synchronize. |
| Communication policy   | None, static, global, entropy, bottleneck | Tests the core safety-control frontier without diluting the study with every policy.         |

**Table 12: Robustness results by regime. Static beacons have the highest information-safety efficiency (ISE) in every hazard-familiarity cell, while the best policy by completed evacuations varies.**

| Hazard | Familiarity | Static ISE | Global/static ISE | Best evacuation count |
|--------|-------------|------------|-------------------|-----------------------|
| Low    | Low         | 0.0155     | 0.69              | No intervention       |
| Low    | Mixed       | 0.0164     | 0.63              | Static beacon         |
| Low    | High        | 0.0144     | 0.68              | Entropy-targeted      |
| Medium | Low         | 0.0126     | 0.44              | Global broadcast      |
| Medium | Mixed       | 0.0130     | 0.44              | Static beacon         |
| Medium | High        | 0.0132     | 0.43              | No intervention       |
| High   | Low         | 0.0098     | 0.29              | Global broadcast      |
| High   | Mixed       | 0.0109     | 0.27              | Global broadcast      |
| High   | High        | 0.0097     | 0.28              | Entropy-targeted      |

**Figure 5: Robustness heatmap from the 900-run grid. Static beacon ISE remains positive across all hazard and familiarity regimes, while the global/static ISE ratio stays below one in every cell. Broad broadcast reach therefore remains less safety-efficient than local static messaging even in regimes where global broadcast wins raw evacuation count.**

### 6.3 Claim Update

The completed robustness study updates the evidence boundary in §5.8. The evidence supports local static messaging as a robust baseline by information-safety efficiency across the tested hazard severity and population familiarity regimes. It does not support the stronger claim that static beacons are the universal operational winner. Evacuation counts remain regime dependent, and the best evacuation-count policy changes across the grid. The correct claim

is therefore narrower and stronger: the information-safety framework reveals a stable local-control advantage that would be obscured by looking only at reach or completed evacuations.

## 7 Related Work

Chiyoda builds on physical pedestrian-evacuation models rather than replacing them. Social-force, cellular-automata, and microscopic interaction models explain how local movement rules can

produce bottlenecks, counterflow, density waves, and other crowd-level failure modes [4, 9, 20]. Evacuation-model reviews caution that such models remain engineering abstractions whose outputs depend on scenario scope, behavioral assumptions, calibration, and validation data [8, 23]. Chiyoda uses this tradition as the motion substrate; its distinction is the information-control layer placed on top of that substrate.

Hazard-coupled evacuation and guidance models are closest to the CBRN-like setting studied here. Toxic-gas and gaseous-material simulations show that contaminant spread, perception, and information transmission can alter route choice and evacuation outcomes [15, 17]. Guidance models similarly show that recommendations must account for hazards, congestion, and compliance [5]. Chiyoda differs by making message policy the controlled variable: timing, targeting, budget, credibility, and reach are varied directly instead of being treated as fixed scenario assumptions.

Incomplete-information and social-behavior models motivate the belief layer. Bayesian and game-theoretic route-choice work studies uncertainty and congestion [10, 29], while social-group ABM surveys and evacuation studies emphasize communication, affiliation, familiar routes, visible crowd movement, and conflicting social cues [2, 12, 14, 25, 27, 28]. Chiyoda keeps these effects stylized, but exports belief entropy, belief accuracy, reach, exposure, queueing, and exit imbalance together so belief improvement can be compared against downstream safety effects.

Existing tools cover much of the surrounding simulation and validation stack. JuPedSim, Vadere, and SUMO provide mature pedestrian simulation capabilities, and PedPy provides trajectory-analysis measures for density, velocity, flow, and fundamental diagrams [1, 11, 13, 24]. Geometry and hazard studies can draw on OpenStationMap, GTFS Pathways, and high-fidelity tools such as the NIST Fire Dynamics Simulator [18, 19, 21]. Chiyoda is therefore best read as a replayable information-intervention benchmark that can be calibrated or cross-checked against these systems, not as a substitute for full pedestrian or dispersion validation [22].

## 8 Limitations

Chiyoda is currently a stylized research simulator, not an operational evacuation planner. The station layouts are simplified grids unless imported from richer geometry. Hazard physics are intentionally lightweight and should not be read as validated CBRN dispersion models. Physiological impairment uses piecewise exposure curves that are useful for comparative simulation but not medical prediction.

The behavioral model is also a hypothesis generator rather than a calibrated description of a specific crowd. Agents use simplified BDI-style intentions, local gossip, panic/anxiety states, and familiarity cohorts. These mechanisms are valuable because they make assumptions inspectable, but the model should not be read as a real-world predictive model without calibration against evacuation drills, incident records, controlled VR experiments, or pedestrian trajectory data. The included Wuppertal bottleneck reference is the first step toward that external calibration: it gives Chiyoda a public trajectory target for bottleneck-flow measurement. It does not by itself calibrate the full behavioral model. The current sweep improves the original proxy but still underestimates flow and overestimates

time headway, so the discrepancy should be read as evidence about the limits of the current grid-scale proxy rather than as a station-level validation result.

Finally, entropy is not equivalent to safety. The central premise is that uncertainty reduction can be harmful when it produces synchronized movement. The proposed information-safety metrics are therefore comparative diagnostics, not universal welfare functions.

## 9 Future Work

The next step is not to claim operational readiness, but to broaden external validity while preserving the safety-control interpretation developed here. Future studies should test richer station layouts, sensor assumptions, hazard models, and calibrated population mixes. They should also compare Chiyoda’s stylized trajectories against evacuation drills, incident records, controlled VR experiments, and pedestrian trajectory datasets before making predictive claims about specific stations or events.

Adaptive message generation is a second extension, including learned or language-model-mediated guidance. Section 5.7 shows that generated guidance can be replayable and safety-efficient under a sparse budget, but it also shows why messages should not be evaluated by fluency alone. Future generated-language systems should remain replayable, validated against the simulated state, prevented from inventing unavailable exits or stale hazard claims, and measured by downstream exposure, queueing, exit balance, and evacuation outcomes.

## 10 Conclusion

Chiyoda reframes evacuation as an information-control problem. In a hazard-coupled crowd, emergency communication changes beliefs, beliefs change routes, and route convergence changes exposure and congestion. The framework therefore measures not only whether information spreads, but whether spreading it at a particular time, place, credibility, and radius improves safety.

The current empirical package supports that framing, but it also narrows the claim. Static beacons are the highest-ISE tested deterministic baseline, global broadcast shows why reach is not enough, and entropy-targeted intervention exposes how belief improvement can coexist with weak aggregate completion. The 900-run robustness grid strengthens this interpretation: static local messaging remains the most efficient information-safety baseline across the tested hazard and familiarity regimes, while the best evacuation-count policy changes by regime.

The LLM extension reinforces the same standard rather than replacing it. Validated sparse generated guidance can be highly information-efficient, and prompt objectives change downstream safety tradeoffs. But equal-budget generated guidance weakens the efficiency advantage, and none of the LLM variants eliminates harmful convergence. This makes generated evacuation guidance scientifically useful as a controlled safety-action proposal, not as an unconstrained source of operational advice.

## References

- [1] Pablo Alvarez Lopez, Michael Behrisch, Laura Bieker-Walz, Jakob Erdmann, Yun-Pang Flötteröd, Robert Hilbrich, Leonhard Lücken, Johannes Rummel, Peter Wagner, and Evamarie Wießner. 2018. Microscopic Traffic Simulation using



- SUMO. In *2018 21st International Conference on Intelligent Transportation Systems*. 2575–2582. doi:10.1109/ITSC.2018.8569938
- [2] Nikolai W. F. Bode and Edward A. Codling. 2013. Human Exit Route Choice in Virtual Crowd Evacuations. *Animal Behaviour* 86, 2 (2013), 347–358. doi:10.1016/j.anbehav.2013.05.025
  - [3] Maik Boltes, Stefan Holl, and Armin Seyfried. 2020. Data Archive for Exploring Pedestrian Dynamics and Its Application in Dimensioning of Facilities for Multi-directional Streams. *Collective Dynamics* 5 (2020), 17–24. doi:10.17815/CD.2020.28
  - [4] Carsten Burstedde, Kai Klauack, Andreas Schadschneider, and Johannes Zittartz. 2001. Simulation of pedestrian dynamics using a two-dimensional cellular automaton. *Physica A: Statistical Mechanics and its Applications* 295, 3–4 (2001), 507–525.
  - [5] James C. Chu, Albert Y. Chen, and Yu-Fu Lin. 2017. Variable Guidance for Pedestrian Evacuation Considering Congestion, Hazard, and Compliance Behavior. *Transportation Research Part C: Emerging Technologies* 85 (2017), 664–683. doi:10.1016/j.trc.2017.10.009
  - [6] John J. Fruin. 1971. *Pedestrian Planning and Design*. Metropolitan Association of Urban Designers and Environmental Planners, New York.
  - [7] Ren-Yong Guo, Hai-Jun Huang, and S. C. Wong. 2012. Route Choice in Pedestrian Evacuation under Conditions of Good and Zero Visibility: Experimental and Simulation Results. *Transportation Research Part B: Methodological* 46, 6 (2012), 669–686. doi:10.1016/j.trb.2012.01.002
  - [8] Steve Gwynne, E. R. Galea, M. Owen, P. J. Lawrence, and L. Filippidis. 1999. A Review of the Methodologies Used in the Computer Simulation of Evacuation from the Built Environment. *Building and Environment* 34, 6 (1999), 741–749. doi:10.1016/S0360-1323(98)00057-2
  - [9] Dirk Helbing and Peter Molnár. 1995. Social force model for pedestrian dynamics. *Physical Review E* 51, 5 (1995), 4282–4286.
  - [10] S. P. Hoogendoorn and P. H. L. Bovy. 2004. Pedestrian Route-Choice and Activity Scheduling Theory and Models. *Transportation Research Part B: Methodological* 38, 2 (2004), 169–190. doi:10.1016/S0191-2615(03)00007-9
  - [11] Arnel Ulrich Kemloh Wagoum, Mohcine Chraïbi, Jun Zhang, and Gregor Lämle. 2015. JuPedSim: An Open Framework for Simulating and Analyzing the Dynamics of Pedestrians. In *Proceedings of the 3rd Conference of Transportation Research Group of India*.
  - [12] Max Kinader, Enrico Ronchi, Daniel Gromer, Mathias Müller, Michael Jost, Markus Nehfischer, Andreas Mühlberger, and Paul Pauli. 2014. Social Influence on Route Choice in a Virtual Reality Tunnel Fire. *Transportation Research Part F: Traffic Psychology and Behaviour* 26 (2014), 116–125. doi:10.1016/j.trf.2014.06.003
  - [13] Benedikt Kleinmeier, Benedikt Zönnchen, Marion Gödel, and Gerta Köster. 2019. VADER: An Open-Source Simulation Framework to Promote Interdisciplinary Understanding. *Collective Dynamics* 4 (2019), 1–34. doi:10.17815/CD.2019.21
  - [14] Margrethe Kobes, Ira Helsloot, Bauke de Vries, and Jos G. Post. 2010. Building Safety and Human Behaviour in Fire: A Literature Review. *Fire Safety Journal* 45, 1 (2010), 1–11. doi:10.1016/j.firesaf.2009.08.005
  - [15] Mengting Liu, Wei Zhu, Yafei Wang, and Jianchun Zheng. 2021. Modeling and Simulation of Exit Selection Behavior in Pedestrian Evacuation Based on Information Perception and Transmission. *Sustainability* 13, 23 (2021), 13194. doi:10.3390/su132313194
  - [16] Hani S. Mahmassani. 1990. Dynamic models of commuter behavior: Experimental investigation and application to the analysis of planned traffic disruptions. *Transportation Research Part A* 24, 6 (1990), 465–484.
  - [17] Jaenudin Makmul. 2020. A Cellular Automaton Model for Pedestrians' Movements Influenced by Gaseous Hazardous Material Spreading. *Modelling and Simulation in Engineering* 2020 (2020), 3402198. doi:10.1155/2020/3402198
  - [18] Kevin B. McGrattan, Randall J. McDermott, Craig Weinschenk, and Glenn P. Forney. 2013. *Fire Dynamics Simulator, Technical Reference Guide, Sixth Edition*. Special Publication 1018. National Institute of Standards and Technology. doi:10.6028/NIST.SP.1018
  - [19] MobilityData. 2026. GTFS Pathways Examples. <https://gtfs.org/documentation/schedule/examples/pathways/>. Accessed 2026-04-28.
  - [20] Mehdi Moussaïd, Dirk Helbing, and Guy Theraulaz. 2011. How Simple Rules Determine Pedestrian Behavior and Crowd Disasters. *Proceedings of the National Academy of Sciences* 108, 17 (2011), 6884–6888. doi:10.1073/pnas.1016507108
  - [21] OpenStreetMap Wiki Contributors. 2026. OpenStreetMap. <https://wiki.openstreetmap.org/wiki/OpenStreetMap>. Accessed 2026-04-28.
  - [22] Enrico Ronchi, Erica D. Kuligowski, Daniel Nilsson, Richard D. Peacock, and Paul A. Reneke. 2013. *The Process of Verification and Validation of Building Fire Evacuation Models*. Technical Note 1822. National Institute of Standards and Technology. doi:10.6028/NIST.TN.1822
  - [23] Andreas Schadschneider, Wolfram Klingsch, Hubert Klüpfel, Tobias Kretz, Christian Rogsch, and Armin Seyfried. 2009. Evacuation Dynamics: Empirical Results, Modeling and Applications. In *Encyclopedia of Complexity and Systems Science*, Robert A. Meyers (Ed.). Springer, New York, 3142–3176. doi:10.1007/978-0-387-30440-3\_187
  - [24] Tobias Schrödter and The PedPy Development Team. 2025. PedPy: Pedestrian Trajectory Analyzer. doi:10.5281/zenodo.17081120
  - [25] Gayani P. D. P. Senanayake, Minh Kieu, Yang Zou, and Kim Dirks. 2024. Agent-based simulation for pedestrian evacuation: A systematic literature review. *International Journal of Disaster Risk Reduction* 111 (2024), 104705. doi:10.1016/j.ijdrr.2024.104705
  - [26] Claude E. Shannon. 1948. A Mathematical Theory of Communication. *Bell System Technical Journal* 27, 3 (1948), 379–423.
  - [27] Jonathan D. Sime. 1983. Affiliative Behaviour during Escape to Building Exits. *Journal of Environmental Psychology* 3, 1 (1983), 21–41. doi:10.1016/S0272-4944(83)80019-X
  - [28] Anne Templeton, Hui Xie, Steve M. V. Gwynne, Aoife Hunt, Pete Thompson, and Gerta Köster. 2024. Agent-based models of social behaviour and communication in evacuations: A systematic review. *Safety Science* 176 (2024), 106520. arXiv:2310.15761 doi:10.1016/j.ssci.2024.106520
  - [29] Y. Wang, E. Ge, and A. Comber. 2023. A Bayesian Nash Equilibrium Model for Pedestrian Evacuation under Incomplete Information. *Journal of Artificial Societies and Social Simulation* 26, 3 (2023).
  - [30] Ulrich Weidmann. 1993. *Transporttechnik der Fussgänger. Schriftenreihe des IVT* 90 (1993).