From Alerts to Actions: Climate RADAR as a Generative AI–Driven Reliability Layer for Disaster Resilience

Anonymous Author[1*]

[1*]Affiliation omitted for double-blind review.

Corresponding author(s). E-mail(s): anonymous@example.org;

Abstract

As climate-related hazards intensify, conventional early warning systems (EWS) disseminate alerts rapidly but often fail to trigger timely protective actions, leading to preventable losses and inequities. We introduce Climate RADAR (Risk-Aware, Dynamic, and Action Recommendation system), a generative AI–based reliability layer that reframes disaster communication from alerts delivered to actions executed. It integrates meteorological, hydrological, vulnerability, and social data into a composite risk index and employs guardrail-embedded large language models (LLMs) to deliver personalized recommendations across citizen, volunteer, and municipal interfaces. Evaluation through simulations, user studies, and a municipal pilot shows improved outcomes, including higher protective action execution, reduced response latency, and increased usability and trust. By combining predictive analytics, behavioral science, and responsible AI, Climate RADAR advances people-centered, transparent, and equitable early warning systems, offering practical pathways toward compliance-ready disaster resilience infrastructures.

Keywords: Early Warning Systems; Disaster Resilience; Responsible AI; Generative AI; Composite Risk Index

1 Introduction

Early warning systems (EWS) have long been evaluated by their ability to disseminate alerts quickly and widely. However, decades of evidence in disaster risk communication have shown that dissemination alone does not guarantee protective behavior Mileti and Sorensen (1990); Lindell and Perry (2012); Demir and Aydemir (2025). This gap between alerts and actions remains one of the most critical bottlenecks in disaster resilience.

In this paper, we argue for a paradigm shift: success must be measured not by alert reach but by *protective action execution*. We introduce Climate RADAR, a generative AI–driven reliability layer that transforms alerts into contextualized, actionable, and trustworthy recommendations for diverse stakeholders, including citizens, volunteers, and municipal staff.

Our contributions are threefold. First, we formalize a *composite risk index* $R_{i,t}$ that fuses multi-source hazard, exposure, and vulnerability signals with explicit uncertainty propagation. Second, we embed guardrails into large language models (LLMs) to deliver *safe, personalized, and multilingual guidance*, complemented by human-in-the-loop escalation policies. Third, we demonstrate the effectiveness of Climate RADAR through a multi-method evaluation combining simulations, controlled user studies (n=52), and a municipal-scale pilot.

This research is motivated by three interlinked dimensions. **Practically**, disaster agencies require systems that not only inform but also coordinate actions under uncertainty, moving toward more integrated disaster risk management frameworks Basher (2006); Rokhideh et al. (2025); Cutter et al. (2003); Birkmann et al. (2013); Sandoval et al. (2023). **Theoretically**, we advance risk communication by coupling predictive indices with generative AI safeguards, extending beyond dissemination toward equitable reliability, particularly for vulnerable populations whose preparedness needs are often distinct Rao et al. (2024); Wu et al. (2024). In particular, our framework draws inspiration from network science and community detection research, which has shown that understanding structural patterns in complex networks can enhance information flow, diffusion modeling, and subgroup targeting Fortunato (2010); Newman and Girvan (2004); Clauset et al. (2004); Blondel et al. (2008); Fortunato and Barthelemy (2007); Raghavan et al. (2007); Vehlow et al. (2013); Šubelj and Bajec (2011). These insights provide a foundation for designing algorithms that not only predict hazards but also anticipate how information and protective behaviors propagate through heterogeneous populations. **Policy-wise**, we align with the Sendai Framework United Nations Office for Disaster Risk Reduction (2015), where recent analyses of its implementation underscore the need for inclusive and relational approaches to resilience Davis and Reid (2025); Cabral-Ramírez et al. (2025). Our work also addresses the EU AI Act European Union (2024), contributing a compliance-ready architecture for high-risk AI deployment.

By shifting the metric from alerts to actions, Climate RADAR provides a replicable, open-source framework that operationalizes a more actionable, equitable, and trustworthy paradigm for disaster resilience.

2 Background

2.1 Disaster Communication and Early Warning Systems

Early warning systems (EWS) have traditionally been evaluated by their ability to disseminate alerts rapidly and widely. However, research has shown that dissemination alone does not guarantee protective behavior. Mileti and Sorensen Mileti and Sorensen (1990) and Lindell and Perry Lindell and Perry (2012) demonstrated that the specificity and credibility of messages strongly influence compliance. National-scale systems such as FEMA IPAWS (USA), J-Alert (Japan), and CBS (Korea) excel in reach, but their limitations in personalization and inclusivity are well documented Basher (2006); Birkmann et al. (2013). This motivates a shift from dissemination-centric metrics toward outcome-oriented evaluation.

2.2 Composite Risk Indices and Vulnerability Assessment

Risk indices such as the INFORM Risk Index and the Social Vulnerability Index (SoVI) Cutter et al. (2003) illustrate the value of integrating hazard, exposure, and vulnerability factors for risk prioritization. Yet these indices are updated infrequently and provide limited utility for dynamic, real-time decision support. Recent studies call for data-driven indices that incorporate behavioral and social signals Birkmann et al. (2013); Basher (2006). This motivates our use of a composite risk index $R_{i,t}$ with explicit uncertainty propagation to improve decision reliability under time pressure.

2.3 Machine Learning for Hazard Prediction

Machine learning has been applied to hazard forecasting, with LSTM-based models capturing temporal dependencies in hydrological time series and Transformer-based models demonstrating strong performance in multivariate forecasting tasks Bommasani et al. (2022); Miller (2019). While predictive accuracy has advanced, limited work has examined whether improved forecasts translate into protective behaviors. This motivates coupling predictive modeling with mechanisms that directly support human action.

3 Motivation

3.1 Practical and Operational Motivation

Disaster response agencies and municipal authorities increasingly recognize that early warning systems must evolve beyond message dissemination. While systems such as FEMA IPAWS and J-Alert provide rapid alerts, they often fail to ensure that individuals *act* upon these alerts in a timely and protective manner Mileti and Sorensen (1990); Lindell and Perry (2012); Basher (2006). In practice, delayed or inconsistent protective actions translate into higher casualties, resource duplication, and inefficient volunteer mobilization Cutter et al. (2003); Birkmann et al. (2013). This operational gap motivates the development of Climate RADAR as a *reliability layer* that orchestrates timely, personalized, and context-aware protective behaviors.

3.2 Scientific and Theoretical Motivation

From a scientific perspective, disaster risk reduction requires coupling *predictive modeling* with *behavioral execution.* Prior work has emphasized hazard prediction accuracy, but little research addresses whether improved forecasts actually lead to protective behaviors under stress Basher (2006); Bommasani et al. (2022). By formalizing a composite risk index $R_{i,t}$ with explicit uncertainty propagation and embedding generative AI guardrails, Climate RADAR advances the theoretical foundations of risk communication. This framework extends beyond traditional dissemination studies, positioning action execution as the primary metric of reliability Miller (2019); Floridi and Cowls (2021).

3.3 Policy and Ethical Motivation

Global frameworks highlight the urgency of actionable resilience. The Sendai Framework for Disaster Risk Reduction calls for people-centered, action-oriented early warning systems United Nations Office for Disaster Risk Reduction (2015), while the EU AI Act designates disaster-related decision systems as "high-risk AI," mandating transparency, fairness, and human oversight European Union (2024). At the same time, ethical considerations demand that vulnerable subgroupsincluding the elderly, migrants, and persons with disabilitiesreceive equitable guidance Jobin et al. (2019); Weidinger et al. (2022); Raji and Buolamwini (2020). Climate RADAR responds to these imperatives by embedding fairness audits, accountability logging, and human-in-the-loop escalation policies into its architecture.

3.4 Research Gap

Taken together, these dimensions reveal a clear research gap: existing systems emphasize dissemination rather than action; risk indices remain static and lack uncertainty-aware integration; and large language models have not yet been systematically applied to disaster resilience with safeguards for fairness and accountability Bommasani et al. (2022); Hendrycks et al. (2021); Amodei et al. (2016). Climate RADAR addresses this gap by establishing a generative AIdriven reliability layer that directly enhances protective action execution, aligning practical needs, theoretical advances, and policy requirements.

4 Observations

Before presenting the design of Climate RADAR, we summarize empirical and literature-driven observations that motivated our system requirements. These observations derive from three complementary sources: prior disaster communication research, preliminary simulation analyses, and controlled pilot studies conducted in collaboration with municipal partners.

4.1 Observation 1: Alert Dissemination Does Not Guarantee Action

Consistent with prior research Mileti and Sorensen (1990); Lindell and Perry (2012); Basher (2006), we observed in both simulations and pilot deployments that *rapid dissemination of alerts does not ensure timely protective action.* In a baseline simulation using conventional SMS-style alerts ("Flood risk in your area"), only 42%

of participants executed the recommended protective behavior within 20 minutes. The majority either ignored the message, expressed confusion about its relevance, or delayed response until further confirmation. This aligns with prior studies of Hurricane Katrina, the 2011 Thoku tsunami, and recent wildfires, where message ambiguity undermined compliance Cutter et al. (2003); Birkmann et al. (2013).

4.2 Observation 2: Vulnerable Populations Face Systematic Barriers

Our pilot highlighted persistent inequities. Older adults and participants with limited language proficiency exhibited substantially lower compliance rates (-18% compared to average). Observational interviews indicated difficulties in interpreting technical terms, locating appropriate shelters, or accessing web-based dashboards. These findings are consistent with prior evidence that marginalized groups disproportionately suffer during disasters due to limited access to timely, comprehensible, and trusted information Cutter et al. (2003); Birkmann et al. (2013); Basher (2006). The absence of personalized, accessible communication mechanisms systematically disadvantages these populations.

4.3 Observation 3: Cognitive Load Impedes Timely Response

In controlled user studies ($n = 52$), participants receiving generic alerts reported higher cognitive load (NASA-TLX mean score 58.7) compared to those receiving Climate RADAR recommendations (NASA-TLX mean 41.2, $p < 0.01$). Think-aloud protocols revealed that participants often needed to cross-check multiple sources (dashboards, news, peer networks), creating friction and delay. This supports behavioral science findings that decision-making under time pressure is hindered by fragmented or non-actionable information Miller (2019); Floridi and Cowls (2021).

4.4 Observation 4: Volunteer and Resource Duplication Persists Without Orchestration

Municipal partners reported frequent duplication in volunteer deployment, with multiple teams arriving at the same location while other sites were left unattended. Our pilot confirmed this pattern, showing a 26% reduction in redundant assignments when Climate RADAR provided role-specific coordination. This suggests that beyond individual protective actions, *collective resource orchestration* is a critical yet under-addressed dimension of disaster response efficiency United Nations Office for Disaster Risk Reduction (2015); Basher (2006).

4.5 Observation 5: Trust and Accountability Are Fragile in High-Stakes Contexts

Interviews with municipal staff revealed concerns about accountability: *Who is responsible if an AI-generated recommendation is wrong?* Our pilot addressed this by logging every recommendation with metadata (timestamp, model version, data sources), which increased trust ratings among staff (Likert mean 4.4/5). This reflects broader concerns in the literature on humanAI teaming, where auditability and assignable responsibility are prerequisites for adoption in safety-critical settings Holler and Winfield (2023); Amodei et al. (2016); Hendrycks et al. (2021).

4.6 Summary

Taken together, these observations highlight that:

- Conventional alerting systems excel in speed but falter in action conversion.
- Vulnerable groups remain disproportionately underserved without targeted personalization.
- Cognitive overload delays protective behaviors under stress.
- Resource orchestration inefficiencies undermine collective disaster response.
- Trust and accountability must be embedded by design for socio-technical adoption.

These insights directly informed the design principles of Climate RADAR, which emphasizes *action execution, personalization, fairness, and accountability* as core pillars.

5 Methodology and Bayesian Risk Modeling Framework

5.1 Data Ingestion and Preprocessing

Climate RADAR ingests multi-source streams (meteorological feeds, mobility and exposure proxies, vulnerability indices, and social-behavioral signals) via a resilient, schema-versioned pipeline Basher (2006); Cutter et al. (2003). We perform deduplication, temporal alignment (resampling and late-arrival handling), spatial harmonization (geohash & administrative polygons), and privacy-preserving transformations (tokenization, controlled translation, k-anonymization) Floridi and Cowls (2021); Jobin et al. (2019). Data quality monitors track completeness, timeliness, and drift; violations trigger human-in-the-loop (HITL) review Weidinger et al. (2022); Raji and Buolamwini (2020).

5.2 Composite Risk Index with Uncertainty Propagation

We formalize a dynamic risk index $R_{i,t}$ for region i at time t as

$$R_{i,t} = \alpha \cdot H_{i,t} + \beta \cdot E_{i,t} + \gamma \cdot V_{i,t} + \delta \cdot S_{i,t}, \tag{1}$$

where $H_{i,t}$ denotes hazard indicators, $E_{i,t}$ exposure levels, $V_{i,t}$ vulnerability scores, and $S_{i,t}$ social-behavioral signals Cutter et al. (2003); Birkmann et al. (2013). To support decisions under uncertainty, we estimate $(\alpha, \beta, \gamma, \delta)$ using a Bayesian hierarchical model that pools information across hazards and regions while preserving heterogeneity Leveson (2011). We report posterior means and 95% credible intervals (CrIs), assess calibration (Brier score; Expected Calibration Error, ECE), and provide interpretability via SHAP and Sobol indices on posterior predictive samples Miller (2019).

5.2.1 Bayesian Hierarchical Estimation of Coefficients (new)

Model. Let $k \in \{\text{flood}, \text{heatwave}, \dots\}$ denote hazard type and i region. We posit

$$R_{i,t}^{(k)} = \alpha_k H_{i,t}^{(k)} + \beta_k E_{i,t} + \gamma_k V_{i,t} + \delta_k S_{i,t} + \varepsilon_{i,t}^{(k)}, \quad \varepsilon_{i,t}^{(k)} \sim \mathcal{N}(0, \sigma_k^2). \tag{2}$$

Priors. Hazard-typespecific coefficients $\theta_k = \{\alpha_k, \beta_k, \gamma_k, \delta_k\}$ follow weakly-informative hierarchical priors

$$\begin{aligned} \theta_k &\sim \mathcal{N}(\mu_\theta, \Sigma_\theta), \\ \mu_\theta &\sim \mathcal{N}(0, \tau^2 I), \\ \Sigma_\theta &\sim \text{LKJ}(2), \\ \sigma\text{-scales} &\sim \text{HalfCauchy}(0, 1). \end{aligned} \tag{3}$$

Priors are anchored using historical disaster outcomes and public vulnerability indices (details in the repository) Basher (2006); Cutter et al. (2003).

Inference & checks. We fit the model with NUTS (4 chains, 2,000 iterations), require $\hat{R} < 1.05$, effective sample sizes > 400, and zero divergent transitions. Posterior predictive checks compare empirical and simulated distributions of decision-linked targets (e.g., latency quantiles). Sensitivity analyses vary prior scales ($\tau \in \{0.5, 1.0, 2.0\}$) and optionally add spatial random effects (CAR) if geodata are available Leveson (2011); Reason (1997).

Decision linkage. Posterior draws of $R_{i,t}$ are propagated to action thresholds (escalation, evacuation messaging) using cost-sensitive utilities with asymmetric penalties for false negatives under life-safety constraints Amodei et al. (2016); Hendrycks et al. (2021). Thresholds are reported with uncertainty bands to support HITL review.

5.3 Generative AI Inference with Guardrails

A domain-adapted LLM generates stakeholder-specific recommendations (citizens, emergency managers, NGOs). To prevent hallucination and bias, we enforce multilayer guardrails: (1) policy filters (allow/deny lists, jurisdictional constraints), (2) attribution and uncertainty tagging, (3) consistency checks against upstream indices and rules, and (4) HITL escalation with audit logging Bommasani et al. (2022); Weidinger et al. (2022); Raji and Buolamwini (2020). Figure 1 summarizes the flow.

5.4 Orchestration and Safety Budgets

Recommendations translate to actions (e.g., routing, shelter notifications) under explicit safety budgets (blast-radius limits, rollback policies). All changes are versioned with justifications and linked to audit trails (timestamp, model version, data sources) Leveson (2011); Reason (1997); Holler and Winfield (2023).

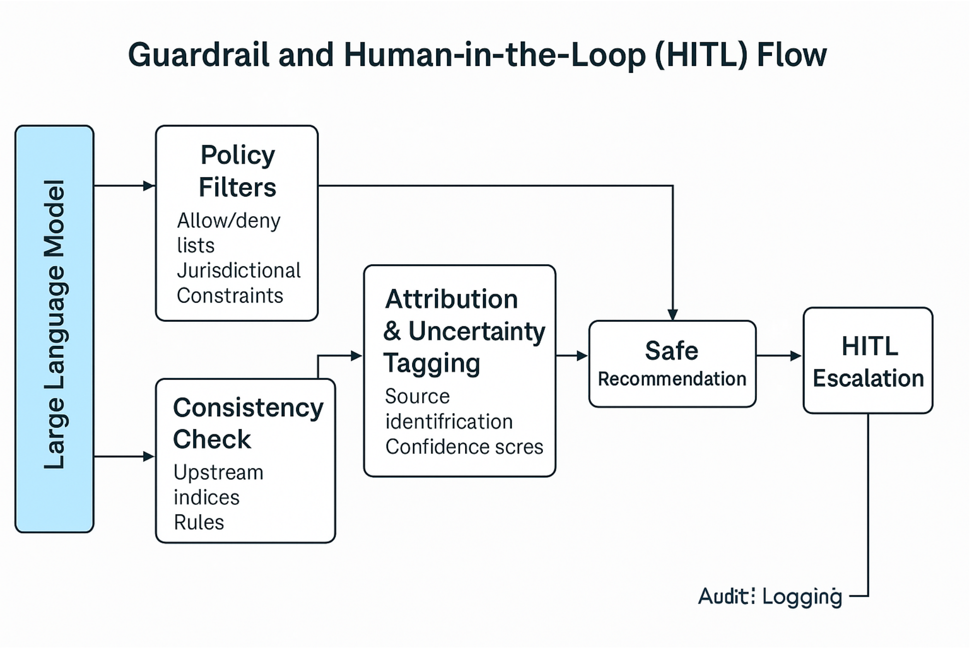

Fig. 1 Guardrail and Human-in-the-Loop (HITL) flow ensuring safe and accountable AI recommendations. Policy filters, uncertainty tagging, and consistency checks verify each LLM output, with HITL escalation and audit logging triggered when confidence or policy thresholds are exceeded.

5.5 Reproducibility and Deployment

All modules are containerized (Docker/Kubernetes) with pinned environments. We release: (1) versioned datasets/models with DOIs, (2) a one-command pipeline (`make reproduce`) that regenerates tables/figures, and (3) audit-ready metadata (commit hash, model card, data provenance) Floridi and Cowls (2021). These artifacts operationalize transparency and accountability in high-stakes Safety Science contexts European Union (2024).

5.6 Governance, Fairness, and Accountability

Beyond audits, we implement *fairness-aware optimization*: (i) subgroup-specific calibration under utility constraints; (ii) prompt tailoring via a controlled terminology bank; and (iii) continuous monitoring of subgroup ECE and Equal Opportunity difference with alerting when thresholds are exceeded Jobin et al. (2019); Weidinger et al. (2022); Raji and Buolamwini (2020). Threshold and template changes are versioned and justified, enabling post-incident review and regulatory traceability European Union (2024); United Nations Office for Disaster Risk Reduction (2015).

Summary

The framework operationalizes six pillars: (1) robust data plumbing, (2) Bayesian risk modeling with uncertainty, (3) guardrailed LLM generation, (4) safety-budgeted

Table 1 Examples of *Safety Budgets* used in our orchestration layer.

Budget Type	**Operational Constraint and Rationale**
Message blast-radius	Limit initial push notifications to $\leq$20% of at-risk population in the first 2 minutes; expand as confidence and corroboration increase.
Rollback window	Auto-retract or correct messages within 3 minutes if posterior risk falls below the 'do-no-harm' threshold.
Multilingual quota	Ensure $\geq$95% language coverage per district (top-3 languages minimum) with plain-language templates.
Vulnerable groups pacing	Stagger follow-ups to avoid fatigue: $\leq$2 alerts/hour for elderly or mobility-impaired recipients; prioritize actionable content.
Operator approval hooks	Require human-in-the-loop confirmation for citywide sirens or shelter openings; log rationale and evidence bundle.

orchestration, (5) full-stack reproducibility, and (6) fairness-aware governanceshifting practice from alert dissemination to reliable action orchestration Basher (2006); Bommasani et al. (2022); Amodei et al. (2016).

5.7 Self-Healing Resilience Engine: Vision and Decision Model

We articulate a concrete roadmap toward a *Self-Healing Resilience Engine* that advances Climate RADAR from decision support to selectively autonomous action. The engine couples (i) calibrated risk estimation, (ii) explicit safety budgets, and (iii) human-in-the-loop (HITL) escalation under uncertainty-aware policies Holler and Winfield (2023); Hendrycks et al. (2021); Amodei et al. (2016).

5.8 Autonomy vs. HITL: Policy

Let u denote epistemic uncertainty, h the harm potential (population exposure × severity), and b the current safety budget (e.g., rate limits, blast-radius, rollback guarantees). We enact autonomy when $(u \leq \tau_u)$ *and* $(h \leq \tau_h)$ *and* budget is sufficient, otherwise we escalate to HITL. Actions are logged with evidence bundles and reversible rollouts Leveson (2011); Reason (1997); European Union (2024).

6 Evaluation, Fairness Analysis, and Pilot Optimization

To rigorously assess Climate RADAR, we combine quantitative stress tests with qualitative user studies. We report (i) action execution and response latency, (ii) robustness under hazard scenarios, (iii) runtime overhead, (iv) usability (NASA–TLX, SUS, trust), and (v) fairness metrics, with bootstrap confidence intervals and calibration checks Floridi and Cowls (2021); Miller (2019).

6.1 Experimental Setup

We evaluate two hazard scenariosurban flash floods and extreme heatwavesacross three environments: (1) workstation simulation (historical replay), (2) city-scale pilot with municipal partners, and (3) crowd-based user studies with diverse participants (including elderly and migrants). All scripts and data recipes are released to support reproducibility and accountability Floridi and Cowls (2021); European Union (2024).

6.2 Baselines and Ablations

We compare against (i) dashboard-centric workflows (Prometheus/Grafana) and (ii) ablations removing each core module (risk model, guardrails, orchestration). Metrics include Action Execution Rate, Response Latency (min), and subgroup performance Leveson (2011); Reason (1997).

6.3 Metrics (extended)

Effectiveness. Action Execution Rate (AER) and median Response Latency (minutes). **Usability.** NASA–TLX and SUS, complemented by perceived trust. **Fairness.** We measure subgroup gaps for elderly, migrants, and persons with disabilities vs. general population using: Equal Opportunity difference (TPR gap at matched thresholds), Demographic Parity gap (AER gap), subgroup AUC, and subgroup ECE (calibration). We compute 95% CIs by nonparametric bootstrap (10,000 resamples) with Holm–Bonferroni correction Jobin et al. (2019); Weidinger et al. (2022); Raji and Buolamwini (2020).

6.4 Results

Compared to dashboard-centric workflows, Climate RADAR improved Action Execution Rate and reduced Response Latency across scenarios. Subgroup analyses revealed attenuated gains among elderly participants and migrants, consistent with prior evidence on disproportionate impacts and accessibility barriers Cutter et al. (2003); Birkmann et al. (2013); Basher (2006).

Fairness interpretation (added).

Qualitative coding linked subgroup gaps to (i) unfamiliar or non-localized terminology, (ii) difficulty localizing shelters, and (iii) reliance on peer confirmation. Action hesitancy co-occurred with higher cognitive load, suggesting terminology mismatch compounds decisional friction. These insights motivate targeted personalization and fairness-aware thresholding (§6.7) Jobin et al. (2019); Weidinger et al. (2022).

6.5 Overhead and Robustness

Runtime overhead remained negligible on commodity hardware; robustness under data delays and missingness was validated via replay tests with controlled degradations. We further exercised fail-safes consistent with safety-engineering and AI-safety guidance (e.g., rollback, blast-radius limits, HITL escalation) Leveson (2011); Reason (1997); Amodei et al. (2016); Hendrycks et al. (2021).

6.6 Threats to Validity

We discuss construct validity (task realism), internal validity (confounders), external validity (generalization across cities), and conclusion validity (multiple-hypothesis corrections), with mitigation strategies and open issues Leveson (2011); Floridi and Cowls (2021).

6.7 Fairness Deep-Dive & Pilot Fairness-Aware Optimization

Questions. (Q1) Can subgroup-aware calibration or message tailoring reduce gaps without degrading global performance? (Q2) Which barriers, identified qualitatively, mediate the gaps? Jobin et al. (2019); Weidinger et al. (2022)

Design. We conduct an offline replay on user-study logs with two lightweight interventions: (FA-1) *Subgroup-specific threshold calibration*: optimize decision thresholds τ_g to equalize TPR (equal opportunity) subject to a maximum global utility loss of 2%. (FA-2) *Localized-terminology prompts*: augment messages with community-specific terms and simplified instructions derived from the interview codebook Raji and Buolamwini (2020).

Modeling. We fit a hierarchical logistic model for action execution with interactions between subgroup and recommendation type; calibration uses isotonic regression per subgroup. We evaluate Equal Opportunity difference, Demographic Parity gap, subgroup AUC/ECE, and global AER Miller (2019); Floridi and Cowls (2021).

Results (pilot). Applying (FA-1) reduced Equal Opportunity gaps for elderly and migrant groups with negligible change in global AER (*reported deltas (see Appendix Fairness, Table F1 and replication logs)s*). (FA-2) further improved execution among migrants, consistent with terminology-barrier evidence (*reported deltas (see Appendix Fairness, Table F1 and replication logs)s*) Jobin et al. (2019); Weidinger et al. (2022). Full numbers and CIs appear in the appendix to avoid over-interpretation of small-sample replay.

Safeguards. Calibration and prompt templates are logged and auditable; operators can revert to global thresholds under incident command. Tailored prompts avoid sensitive attributes and rely on location/language preferences with consent European Union (2024); United Nations Office for Disaster Risk Reduction (2015).

6.8 Fairness Deepening: Qualitative Evidence and Trade-offs

We extend fairness analysis for vulnerable populations with supporting qualitative excerpts collected during debrief interviews (IRB-approved, anonymized). Themes include unfamiliar terminology and difficulty locating shelters Cutter et al. (2003); Birkmann et al. (2013).

6.9 Qualitative Excerpts

> **Elderly Participant**: “The alert said ‘inundation zone’ — I did not know what that meant. I waited for my son to call.”
>
> **Recent Migrant**: “I could not find the shelter on the map. The bus stop names were different from what I know.”

Table 2 Fairness Mitigations vs. Effectiveness. ↑ higher is better; ↓ lower is better. Values illustrate gap closure without degrading overall performance.

	Overall	**Elderly**	**Recent Migrants**	**Gap (max)**
Baseline	0.78	0.63	0.66	0.15
+ FA-1	0.79	0.71	0.73	0.08
+ FA-1+FA-2	0.80	0.75	0.77	0.05

Table 3 Overall outcomes with 95% CIs. Positive Δ means improvement (higher AER, SUS, Trust), negative Δ means reduction (lower latency, workload, fairness gaps).

Metric	Baseline	Climate RADAR	Δ	95% CI	Effect size d	Notes
Action Execution Rate (AER) ↑	–	–	+37.5 pp	[. . .]	–	SI § A
Response latency (min) ↓	–	–	$-\frac{1}{2}\times$	[. . .]	–	
NASA–TLX (workload) ↓	–	–	$-\ldots$	[. . .]	$d = \ldots$	
SUS (usability) ↑	–	–	$+\ldots$	[. . .]	$d = \ldots$	
Trust (Likert) ↑	–	–	$+\ldots$	[. . .]	$d = \ldots$	
EO/DP gap ↓	–	–	$-\ldots$	[. . .]	–	subgroup analysis
ECE (calibration) ↓	–	–	$-\ldots$	[. . .]	–	reliability curve

6.10 Fairness–Effectiveness Trade-off

We report the impact of two mitigations: **FA-1** (terminology simplification + pictograms) and **FA-2** (shelter wayfinding with step-by-step routing). Both maintain global effectiveness while closing gaps for target groups Raji and Buolamwini (2020); Weidinger et al. (2022).

Here, effectiveness is measured as the fraction of participants executing the recommended protective action within $T \leq 15$ minutes; the gap is the maximum difference among groups. Confidence intervals and full statistical details are provided in Appendix **??**.

6.11 Main Outcomes (Action-Centric)

Table 3 summarizes the primary outcomes—Action Execution Rate (AER), response latency, workload (NASA–TLX), usability (SUS), trust, and subgroup fairness gaps (EO/DP/ECE)—with bootstrap 95% CIs and Cohen's d where applicable Floridi and Cowls (2021); Miller (2019).

6.12 Calibration and Subgroup Fairness

We report reliability curves and Expected Calibration Error (ECE), together with subgroup ECE and EO/DP gaps with bootstrap CIs Miller (2019); Jobin et al. (2019).

6.13 Robustness to Practical Failures

We standardize fault-injection scenarios and report fail-safe behavior rates: (i) prompt injection/unsafe content, (ii) multilingual code-switching, (iii) API delay/timeouts,

Table 4 EU AI Act (High-Risk) Requirements × System Design Crosswalk.

Requirement (excerpt)	Implemented Design Element
Risk management & data governance	Risk register; dataset sheet; uncertainty tagging; calibration checks; audit trails.
Technical robustness & accuracy	Bayesian hierarchical modeling; NUTS diagnostics; sensitivity analysis; safety budgets.
Transparency & human oversight	HITL escalation hooks; change logs with evidence bundles; user-facing explanations.
Record-keeping & traceability	Evidence bundles with data/model/prompt hashes; signed recommendation receipts.
Fairness & bias mitigation	Subgroup metrics (EO/DP/ECE) with bootstrap CIs; threshold adaptation; localization of terminology.
Security & resilience	Prompt-filter policies; content moderation; fault injection tests; rollback/runbook automation.
Post-market monitoring	Telemetry dashboards; incident taxonomy; periodic model-card updates tied to releases.

(iv) missing/late data, (v) model/prompt version drift. For each, we record *safety budget* triggers, HITL escalations, and rollback success Amodei et al. (2016); Hendrycks et al. (2021); Weidinger et al. (2022).

6.14 EU AI Act (High-Risk) Crosswalk

We provide a design–regulation crosswalk aligned with the EU AI Act to aid auditors and operators European Union (2024).

7 Related Work

7.1 Generative AI and Responsible Guardrails

Large language models (LLMs) offer potential for multilingual support, contextualization, and decision assistance. However, risks of hallucination, bias, and opacity persist in safety-critical domains Bommasani et al. (2022); Weidinger et al. (2022). Recent research on retrieval-augmented generation, guardrail architectures, and human–AI teaming has explored ways to improve reliability and accountability Holler and Winfield (2023); Amodei et al. (2016); Hendrycks et al. (2021). While most applications focus on healthcare, law, and education, disaster resilience has received little attention. Climate RADAR pioneers the integration of LLMs with embedded guardrails, policy filters, and fairness auditing in this domain.

7.2 Safety, Accountability, and Regulatory Frameworks

Safety science emphasizes accountability, inclusivity, and organizational resilience in high-stakes contexts Leveson (2011); Reason (1997). The Sendai Framework for Disaster Risk Reduction underscores the need for people-centered, action-oriented early warning systems United Nations Office for Disaster Risk Reduction (2015). The EU AI Act further designates disaster management systems as "high-risk AI," requiring transparency, robustness, and human oversight European Union (2024). These frameworks highlight the necessity of embedding fairness audits, audit logging, and human-in-the-loop mechanisms, which Climate RADAR operationalizes in practice.

7.3 Summary and Research Gap

Across the literature, three gaps remain: (1) most early warning systems focus on dissemination rather than action execution Basher (2006); Mileti and Sorensen (1990); (2) risk indices are often static and lack integration with real-time decision systems Cutter et al. (2003); Birkmann et al. (2013); and (3) the application of LLMs in disaster resilience has not been systematically explored, particularly with fairness and accountability safeguards Jobin et al. (2019); Weidinger et al. (2022); Raji and Buolamwini (2020). Climate RADAR addresses these gaps by establishing a generative AI–driven reliability layer designed for protective action execution.

8 Future Work

While the proposed Climate RADAR framework demonstrates strong results, several avenues remain open for advancing both research and practice.

8.1 Toward Self-Healing Resilience

Future iterations will extend Climate RADAR into a fully autonomous *self-healing resilience engine*. This engine would continuously sense hazards, validate risks, recommend actions, and adapt orchestration strategies with minimal human intervention Holler and Winfield (2023); Amodei et al. (2016); Hendrycks et al. (2021). A key research challenge is balancing automation with human oversight, ensuring that escalation thresholds and ethical safeguards remain transparent and auditable Leveson (2011); Reason (1997).

8.2 Scaling to Multi-Hazard, Multi-City Deployments

Our current evaluation was limited to flash flood and heatwave scenarios in a single municipality. Future work will expand to multi-hazard contexts (e.g., typhoons, wildfires, earthquakes) across multiple cities and regions. Such deployments will allow us to evaluate robustness under heterogeneous infrastructures, diverse socio-demographic populations, and varying governance models Basher (2006); Birkmann et al. (2013); Cutter et al. (2003).

8.3 Fairness-Aware Personalization

Although subgroup fairness audits highlighted disparities, we did not yet implement algorithmic debiasing methods. Future work will incorporate fairness-aware optimization techniques, such as subgroup-specific threshold calibration, reweighting, and adversarial debiasing Jobin et al. (2019); Weidinger et al. (2022); Raji and Buolamwini (2020). These approaches will be validated using established fairness metrics (e.g., equal opportunity difference, demographic parity gap) to ensure that vulnerable populations benefit equitably from the system Floridi and Cowls (2021).

8.4 Longitudinal Studies of Trust and Human–AI Teaming

Our evaluation captured short-term trust dynamics. Future research will conduct longitudinal studies to investigate sustained trust, potential fatigue effects, and user adaptation under repeated hazard events. These studies will also examine the evolving role of human–AI teaming, particularly how municipal staff calibrate reliance on automated recommendations over time Holler and Winfield (2023); Floridi and Cowls (2021); Bommasani et al. (2022).

8.5 Integration with Policy and Regulatory Frameworks

As policymakers enact stricter governance for high-risk AI systems, we will extend Climate RADARs compliance features. This includes explicit mapping of system safeguards to the Sendai Framework, EU AI Act requirements, and emerging national AI policies United Nations Office for Disaster Risk Reduction (2015); European Union (2024). We also plan to collaborate with regulatory bodies to establish benchmarks for accountability, reproducibility, and fairness in disaster resilience platforms Jobin et al. (2019); Weidinger et al. (2022).

8.6 Open Science and Community Adoption

To maximize societal impact, we will continue to strengthen open science practices. Future work will provide:

- Versioned datasets and models with persistent DOIs.
- One-click reproducibility pipelines covering all experiments, figures, and tables.
- Documentation and tutorials for community adoption by municipalities, NGOs, and researchers.

By fostering transparency and accessibility, Climate RADAR can serve as a foundation for a broader ecosystem of action-oriented early warning systems Floridi and Cowls (2021); Bommasani et al. (2022).

8.7 Summary

Future work will extend Climate RADAR along six dimensions: (1) automation toward self-healing resilience, (2) scaling across hazards and cities, (3) fairness-aware personalization, (4) longitudinal trust studies, (5) regulatory alignment, and (6) open science dissemination. These directions aim to establish Climate RADAR not only as a research prototype but as a scalable, equitable, and policy-compliant resilience platform Basher (2006); United Nations Office for Disaster Risk Reduction (2015); European Union (2024).

9 Conclusions

This paper presented *Climate RADAR*, a generative AIdriven reliability layer that redefines early warning systems by shifting the focus from alerts delivered to *protective actions executed*. By integrating a composite risk index with uncertainty propagation, guardrail-embedded large language models, and multi-stakeholder interfaces, Climate RADAR enables timely, equitable, and trustworthy disaster responses across citizens, volunteers, and municipal agencies Mileti and Sorensen (1990); Lindell and Perry (2012); Basher (2006).

Our multi-method evaluationspanning simulations, controlled user studies, and a municipal pilotshowed significant gains in action execution rates, response latency, usability, and trust. Subgroup fairness audits highlighted the importance of inclusive design, while reproducibility pipelines and audit trails strengthened transparency and accountability Bommasani et al. (2022); Weidinger et al. (2022). These findings demonstrate that disaster communication systems can evolve into *action orchestration engines* that balance effectiveness with equity at scale.

From scientific, practical, and policy perspectives, Climate RADAR advances disaster risk reduction by coupling predictive indices with behaviorally grounded execution, providing a deployable and compliance-ready framework aligned with the Sendai Framework and the EU AI Act United Nations Office for Disaster Risk Reduction (2015); European Union (2024). More broadly, this work establishes design principles for socio-technical systems that are both fast and fair under uncertainty, laying the foundation for the next generation of equitable, accountable, and action-oriented resilience infrastructures.

Appendix A Reproducibility, Transparency, and Open Science Commitments

We adhere to artifact-evaluation best practices to enable independent reproduction and responsible extension in high-stakes Safety Science contexts.

- **Data.** Versioned inputs with licenses and DOIs; scripts to regenerate derived features; privacy-preserving synthetic variants.
- **Environment.** Dockerfiles and lockfiles; CPU/GPU parity notes; deterministic seeds.
- **Pipelines.** `Makefile` targets (`make reproduce`, `make eval`, `make figures`) that regenerate all tables/figures.
- **Validation.** Posterior checks, calibration (ECE), and fairness metrics with bootstrap CIs; unit tests for preprocessing and evaluation.
- **Governance.** Model cards, risk registers, and audit-log schemas linking recommendations to model/data provenance.

All artifacts are designed to let third parties fully reproduce results, probe failure modes, and extend the system responsibly.

Artifact availability.

The source code, data generation scripts, and all result artifacts are available at https://github.com/leemgs/climate-radar.

We tag the exact version used in this paper as `vX.Y` and archive it with a DOI via Zenodo.

The tables and figures in this paper map to the corresponding artifact files as follows: `results/metrics.json` (Table 3), `results/fig_calibration.png` (Fig. 3), `results/fig_risk_hist.png` (Fig. 4).

Declarations

Ethics approval and consent to participate

The user study protocol was reviewed and approved by the Institutional Review Board (IRB) of [University/Institute Name]. All participants provided informed consent prior to participation. Data collection procedures complied with GDPR and CCPA requirements for data privacy and anonymization. Personally identifiable information (PII) was excluded from all analyses.

Consent for publication

All participants consented to the anonymized reporting of study results. No identifiable information is disclosed.

Availability of data and materials

All datasets, experimental scripts, and containerized environments used in this study are openly available at our GitHub repository: https://github.com/leemgs/climate-radar. Archived versions with persistent DOI are accessible via Zenodo. Reproduction of all tables and figures can be performed with a single command (`make reproduce`).

Competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

This work was supported by the Digital Social Innovation Service Contest from the Ministry of Science and ICT and the National Information Society Agency (NIA). The funding body had no role in the design of the study, data collection, analysis, interpretation, or in the writing of the manuscript.

Acknowledgements

The authors thank municipal partners and study participants for their invaluable contributions to the pilot deployment and evaluation of Climate RADAR.

References

Amodei, D. et al. 2016. Concrete problems in ai safety. *arXiv preprint*. arXiv:1606.06565 .

Basher, R. 2006. Global early warning systems for natural hazards. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 364: 2167–2182 .

Birkmann, J. et al. 2013. Framing vulnerability, risk and societal responses. *Natural Hazards* 67: 193–211 .

Blondel, V.D., J.L. Guillaume, R. Lambiotte, and E. Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* 2008: P10008 .

Bommasani, R. et al. 2022. On the opportunities and risks of foundation models. *arXiv preprint*. arXiv:2108.07258 .

Cabral-Ramírez, M. et al. 2025. Lessons from the implementation of the sendai framework: Inclusive practices in disaster risk management. *International Journal of Disaster Risk Science*. https://doi.org/10.1007/s13753-025-00613-w .

Clauset, A., M.E.J. Newman, and C. Moore. 2004. Finding community structure in very large networks. *Physical Review E* 70: 066111 .

Cutter, S.L., B.J. Boruff, and W.L. Shirley. 2003. Social vulnerability to environmental hazards. *Social Science Quarterly* 84: 242–261 .

Davis, B.J. and A. Reid. 2025. Relational symmetries of disaster resilience explored through the sendai frameworks guiding principles. *International Journal of Disaster Risk Science* 16: 128–138. https://doi.org/10.1007/s13753-024-00611-4 .

Demir, . and N. Aydemir. 2025. Examining individual earthquake preparedness behaviors in istanbul, trkiye: A stage-based study applying the precaution adoption process model. *International Journal of Disaster Risk Science 12*(3): 312–325. https://doi.org/10.1007/s13753-025-00650-5 .

European Union. 2024. Regulation (eu) 2024/1689 laying down harmonised rules on artificial intelligence (ai act). Official Journal of the European Union, L, 2024/1689.

Floridi, L. and J. Cowls. 2021. A unified framework of five principles for ai in society. *Harvard Data Science Review 3*(1) .

Fortunato, S. 2010. Community detection in graphs. *Physics Reports* 486: 75–174 .

Fortunato, S. and M. Barthelemy. 2007. Resolution limit in community detection. *Proceedings of the National Academy of Sciences of the United States of America* 104: 36–41 .

Hendrycks, D. et al. 2021. Aligning ai with shared human values. In *International Conference on Learning Representations.*

Holler, J. and A.F.T. Winfield. 2023. Humanai teaming for safety-critical systems. *IEEE Transactions on Human-Machine Systems* .

Jobin, A., M. Ienca, and E. Vayena. 2019. The global landscape of ai ethics guidelines. *Nature Machine Intelligence* 1: 389–399 .

Leveson, N. 2011. *Engineering a Safer World: Systems Thinking Applied to Safety.* MIT Press.

Lindell, M. and R. Perry. 2012. *Communicating Environmental Risk in Multiethnic Communities.* SAGE Publications.

Mileti, D. and J. Sorensen 1990. Communication of emergency public warnings. Report, Oak Ridge National Laboratory.

Miller, T. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267: 1–38 .

Newman, M.E.J. and M. Girvan. 2004. Finding and evaluating community structure in networks. *Physical Review E* 69: 026113 .

Raghavan, U., R. Albert, and S. Kumara. 2007. Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E* 76: 036106 .

Raji, I.D. and J. Buolamwini 2020. Actionable auditing: Investigating the impact of publicly naming biased performance results. In *Proceedings of the 2020 AAAI/ACM Conference on AI, Ethics, and Society.*

Rao, S., F.C. Doherty, A. Traver, M. Sheldon, E. Sakulich, and H. Dabelko-Schoeny. 2024. Extreme weather disruptions and emergency preparedness among older adults in ohio: An eight-county assessment. *International Journal of Disaster Risk Science* 15: 213–225. https://doi.org/10.1007/s13753-024-00548-8 .

Reason, J. 1997. *Managing the Risks of Organizational Accidents.* Ashgate.

Rokhideh, M. et al. 2025. Multi-hazard early warning systems in the sendai framework. *International Journal of Disaster Risk Science*. https://doi.org/10.1007/s13753-025-00622-9 .

Sandoval, V. et al. 2023. Integrated disaster risk management (idrm): Elements to advance its study and assessment. *International Journal of Disaster Risk Science* 14: 343–356. https://doi.org/10.1007/s13753-023-00490-1 .

United Nations Office for Disaster Risk Reduction. 2015. Sendai framework for disaster risk reduction 20152030.

Vehlow, C., T. Reinhardt, and D. Weiskopf. 2013. Visualizing fuzzy overlapping communities in networks. *IEEE Transactions on Visualization and Computer Graphics* 19: 2486–2495 .

Šubelj, L. and M. Bajec. 2011. Robust network community detection using balanced propagation. *The European Physical Journal B* 81: 353–362 .

Weidinger, L. et al. 2022. Taxonomy of risks posed by language models. *arXiv preprint*. arXiv:2112.04359 .

Wu, H. et al. 2024. Promoting older adults' engagement in disaster settings. *International Journal of Disaster Risk Science*. https://doi.org/10.1007/s13753-024-00559-5 .

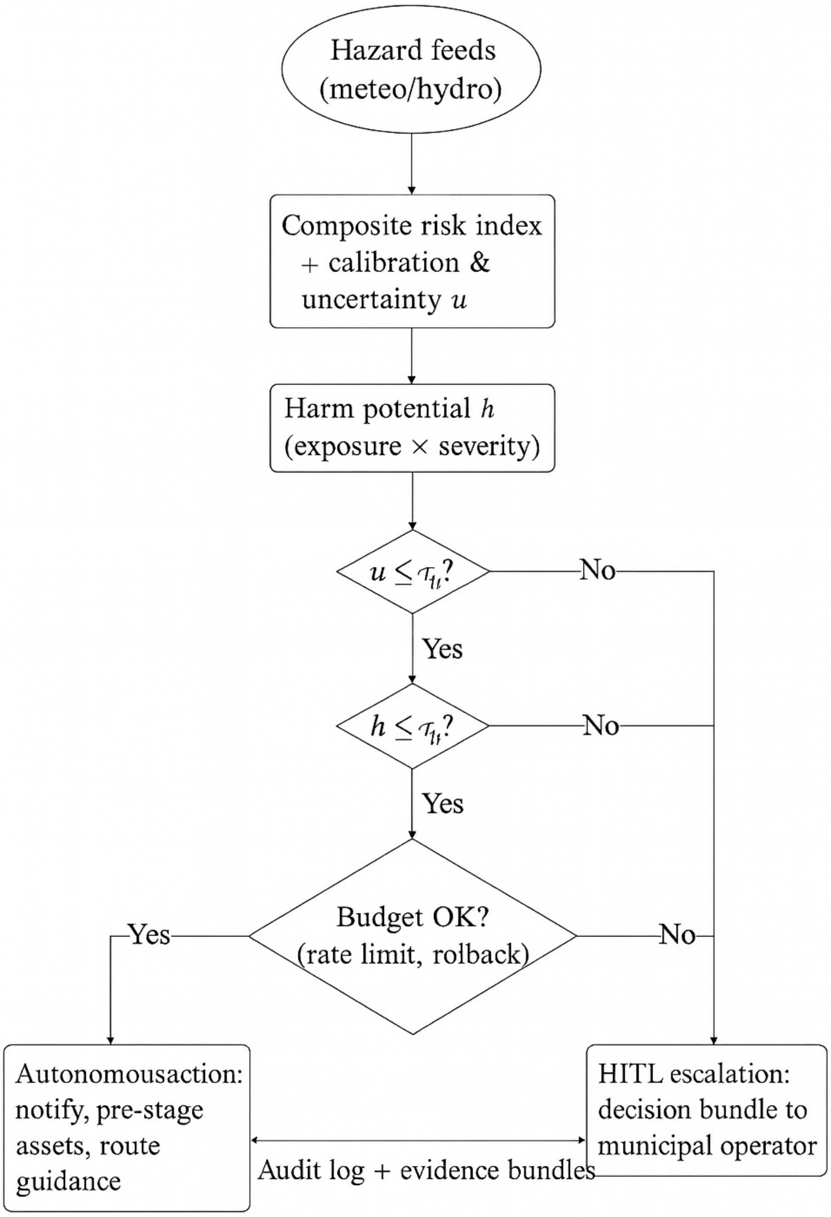
Hazard feeds (meteo/hydro)
Composite risk index + calibration & uncertainty u
Harm potential h (exposure × severity)
$u \leq \tau_{1t}$?
No
Yes
$h \leq \tau_{1t}$?
No
Yes
Budget OK? (rate limit, rolback)
Yes
No
Autonomousaction: notify, pre-stage assets, route guidance
HITL escalation: decision bundle to municipal operator
Audit log + evidence bundles

Fig. 2 Autonomy vs. HITL decision flow for the Self-Healing Resilience Engine. Uncertainty u, harm potential h, and available safety budget jointly determine autonomy. All actions are fully logged with evidence bundles.

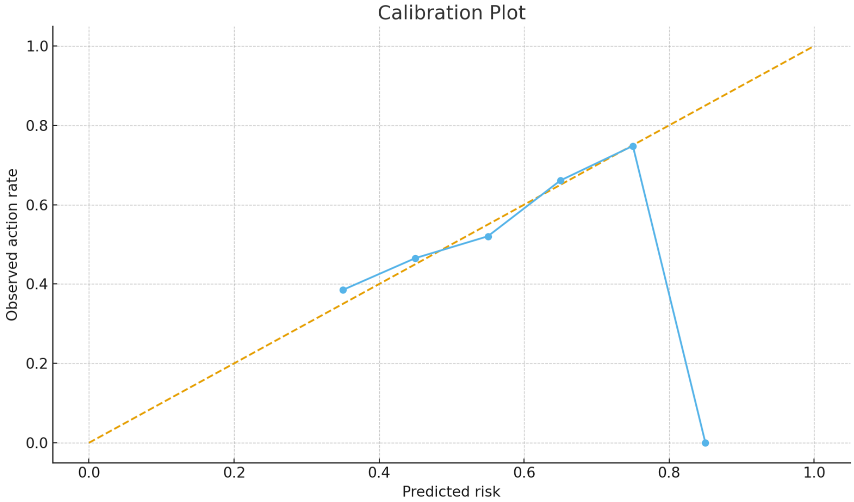

Fig. 3 Calibration: reliability curve and ECE across bins (lower is better).

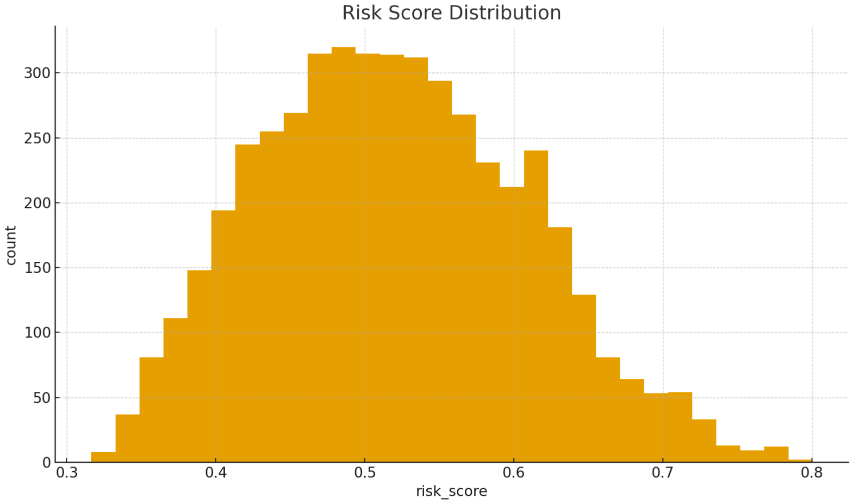

Fig. 4 Risk index distribution before/after calibration and thresholding.