Simulating Crisis Cognition: A Computational Framework for Hypothesis Generation in Crisis Communication

Changsen Yuan[1,3], Yanghao Zhou[2,3], Chong Feng[2], Ge Shi[2]*
[1]Beijing University of Technology, Beijing, China
[2]Beijing Institute of Technology, Beijing, China
[3]Beijing Institute of Technology Southeast Academy of Information Technology, Fujian, China
yuanchangsen@bjut.edu.cn, {zhouyh77,fengchong,gshi}@bit.edu.cn

Abstract

Large Language Models (LLMs) have demonstrated remarkable fidelity in simulating social dynamics, yet using them to inform high-stakes crisis policy requires rigorous causal evaluation. We introduce CRISIS COGNITION, a framework rooted in generative Structural Causal Models (SCM) that functions as an *in-silico* hypothesis generator. By coupling real-world telemetry with 1,813 agents, we conduct a counterfactual simulation to evaluate communication strategies. **Unlike prior descriptive work, we employ a Stratified Analysis to strictly control for personality confounders.** Our simulations generate a **computational hypothesis**: within the LLM's generative process, emotional scaffolding serves as a functional prerequisite to unlock valid reasoning paths for *high-neuroticism* agents. Crucially, we identify a "Sedative Effect" in simultaneous interventions, confirming that the *sequence* of support is as vital as the content. This framework provides a rigorous testbed for evaluating strategies before human-subject trials.

1 Introduction

The integration of Large Language Models (LLMs) into Computational Social Science (CSS) has facilitated the growth of **"Generative Social Simulation"** (Park et al., 2023; Gao et al., 2023; Argyle et al., 2022), where agents act as sophisticated simulacra of human behavior. However, much of this work remains predominantly *descriptive*, focusing on *how* phenomena like misinformation spread or panic emerge, while stopping short of providing *prescriptive* insights. Knowing mechanics of chaos does not inherently inform systems *how* to generate optimal textual interventions to mitigate it.

This descriptive-predictive focus leaves a critical **prescriptive gap** in high-stakes domains like crisis management. Policymakers cannot ethically perform Randomized Controlled Trials (RCTs) during disasters to test, for instance, whether an empathetic message is effectively superior to an authoritative one. This **"Counterfactual Gap"** (see Figure 1) forces a reliance on automated systems built on rule-based heuristics or standard Retrieval-Augmented Generation (RAG) (Lewis et al., 2020; Shuster et al., 2021; Jeong et al., 2024). These baselines often prioritize factual accuracy and information completeness while underestimating the **psycholinguistic receptivity** of the message under acute stress, a factor long emphasized in classical stress appraisal theories (Lazarus and Folkman, 2020).

We bridge this gap by reframing crisis communication simulation through a **Generative Structural Causal** lens. We propose CRISIS COGNITION, a framework that leverages LLMs not merely as text generators, but as an *in-silico* counterfactual laboratory. By formalizing the crisis interaction within a Generative SCM $\mathcal{M} = \langle U, E, T, Y \rangle$, we apply Pearl's *do*-calculus (Zelterman, 2001; Feder et al., 2022; Gkountouras et al., 2025) to **isolate the effect** of linguistic strategies (T). **Crucially, because we possess full control over the inputs (U, E), the causal effects are identifiable by design.** Our proposed CRISISCOGNITION framework, illustrated in Figure 1.

A key technical contribution is the construction of **COPE-CF** (Crisis Offline-Online Parallel Evaluation - Counterfactual). Unlike observational datasets where user traits (U) and treatments (T) are correlated, COPE-CF leverages the simulation's unique capability to **hold the agent's persona and environment constant** while intervening solely on the message ($do(T)$). This creates **perfect counterfactual pairs** impossible to obtain in human-subject research, enabling precise estimation of the treatment's marginal effect within the simulation. We **emphasize** that our findings constitute *computational hypotheses* derived from LLM-based simulations. They are intended to serve as a

* Corresponding author

Findings of the Association for Computational Linguistics: ACL 2026, pages 14140–14148

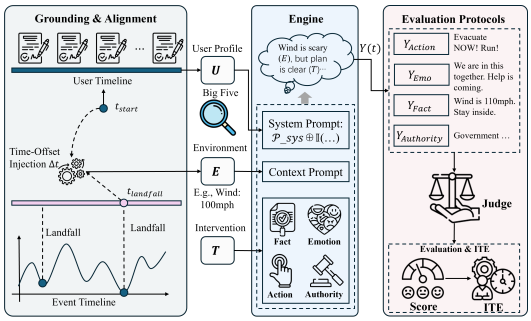

Figure 1: The CRISIS COGNITION Framework. The pipeline consists of: (Left) Temporal-Semantic Alignment to derive environment $\mathbf{E}$ and personality $\mathbf{U}$; (Center) Cognitive Execution Engine using parametric prompting to form cognitive state $\mathbf{C}$; (Right) Evaluation Protocols where a Judge Agent quantifies outcomes $\mathbf{Y}$ to compute ITE.

theoretically-grounded blueprint for guiding future human-subject research, rather than as immediate policy prescriptions.

Our contributions are threefold: (1) CRISIS COGNITION, a framework for generating hypothesis-driven simulations; (2) **COPE-CF**, a benchmark dataset containing **paired counterfactual trajectories** that rigorously disentangle personality form intervention effects; and (3) computational evidence suggesting that psychological safety is a functional prerequisite for effective information processing.

2 Methodology

Our objective is to transition crisis simulation from descriptive mimicry to prescriptive estimation within a controlled computational environment.

2.1 Generative Structural Causal Formulation

We formalize the crisis interaction as a Generative SCM $\mathcal{M} = \langle U, E, T, Y \rangle$ (Zelterman, 2001), where U denotes agent persona (System Prompt), E denotes environment (Context), T is the intervention, and Y is the outcome.

Identifiability by Design. A core critique of applying SCM to LLMs is the "Black Box" nature of neural networks. However, in our simulation, the confounders are not latent. U and E are **explicit inputs** fully observable and controlled by the experimenter. Crucially, the treatment assignment T is randomized independently of U ($T \perp U$). Therefore, the **Backdoor Criterion** is satisfied *by design* (see in Appendix C), allowing us to identify the causal effect $P(Y|do(T))$ directly as $P(Y|T, U, E)$ without unobserved confounding bias.

Backdoor Adjustment via Stratification. To estimate the Individual Treatment Effect (ITE), we move beyond simple linear regression. Instead, we perform a **Stratified Analysis**, partitioning agents into subgroups based on Neuroticism scores (U). This allows us to observe heterogeneous treatment effects without assuming a linear relationship between personality and anxiety.

2.2 Layer 1: Grounding and Alignment

To prevent "contextual hallucination" common in open-ended LLM simulations, we construct the **COPE-CF** (Crisis Offline-Online Parallel Evaluation - Counterfactual) benchmark to anchor agents in objective reality.

Personality Encoding. We initialize agents using the validated PSYCHOAGENT dataset (Liu et al., 2025), leveraging its pre-computed personality traits and historical telemetry alignment. For the Neuroticism trait, ϕ achieved a Pearson correlation of $r = 0.76$ ($p < .001$). Recognizing this as a **noisy measurement**, we treat the inferred scores as continuous variables rather than binary categories to minimize information loss.

Temporal-Semantic Alignment. We synchronize agent timelines with historical hurricane peak scenarios using a deterministic algorithm. At each step, physical telemetry is discretized into a standardized template: *"Current Environment: Wind Speed {wind_speed} mph, Pressure {pressure} mb."* **This context is injected directly into the system prompt**, ensuring that the environmental severity E is the primary exogenous driver of the agent's situational awareness.

2.3 Layer 2: Cognitive Execution Engine

The execution layer maps SCM nodes to modular prompt components, utilizing a **CoT** mechanism to simulate human-like risk processing.

Cognitive Workflow. The engine operates via a two-stage prompt: (1) *Appraisal Phase*: The agent generates a CoT log C assessing personal risk; (2) *Response Phase*: The agent maps C to a behavioral reply Y. We operationalize the cognitive mediator C by parsing the CoT logs. **Threat-appraisal tokens** are identified via a sentiment-aware lexicon focused on risk, while **planning-oriented tokens** are identified via imperative survival verbs.

Method Strategy	GPT-4o		DeepSeek-V3		Llama-3.1-70B	
	Anxiety (Y) ↓	Action (A) ↑	Anxiety (Y) ↓	Action (A) ↑	Anxiety (Y) ↓	Action (A) ↑
RAG	5.75 ±1.2	0.75 ±0.4	7.75 ±0.9	0.46 ±0.3	6.43 ±1.1	0.69 ±0.5
SIMULTANEOUS	4.83 ±1.0	0.10 ±0.1	5.82 ±1.3	0.05 ±0.1	3.30 ±0.8	0.18 ±0.2
SEQUENTIAL	**3.57** ±0.9	**4.47** ±1.5	7.17 ±1.4	**4.83** ±1.8	3.68 ±0.7	**5.74** ±1.6

Table 1: **Stratified Efficacy Analysis across Backbones.**

Mechanistic Interpretability via CoT. To provide process-level evidence for the "Cognitive Bottleneck," we introduce an intermediate cognitive state C. Specifically, the $Prompt$ directs the LLM to reason step-by-step. This allows us to observe whether T successfully penetrates the agent's internal reasoning before it manifests in behavior Y.

Decoupled Outcome Evaluation. To eliminate **self-preference bias**, we strictly decouple simulation from evaluation. While agents are instantiated via DeepSeek-V3 (DeepSeek-AI et al., 2024) or Llama-3.1-70B (Dubey et al., 2024), the behavioral outcomes (Y) and anxiety levels are quantified by a blinded **GPT-4o judge**. This was calibrated against human-annotated samples (Pearson $r = 0.82$) and remained unaware of treatment condition.

2.4 Layer 3: Evaluation Protocols

To rigorously assess the causal impact of linguistic interventions, we established a tri-dimensional evaluation framework covering behavioral outcomes, cognitive processes, and metric validity.

Primary Outcome: Behavioral Agency (A). Our core metric, *Valid Action Density*, quantifies the agent's functional capacity to execute survival protocols. **(1) Operationalization:** We employ a rule-based parser that counts unique survival-oriented verbs (e.g., *evacuate, seal*) mapped to the FEMA resilience taxonomy. **(2) Negation Filtering:** To ensure construct validity, we implemented a **Negation-Filtered Heuristic**. Verbs modified by negative particles (e.g., "I *cannot* evacuate") are explicitly excluded. This keeps false positives where agents express helplessness rather than agency.

Secondary Outcome: Psychological State (Y). We quantify *Anxiety Score* (Y) on a Likert scale (1-10). To eliminate self-preference bias, outcomes are evaluated by an independent, **blinded LLM judge** (GPT-4o) unaware of treatment condition.

Mechanistic Metric: Cognitive Bandwidth ($\mathcal{B}$). To explain the "Cognitive Bottleneck," we operationalize bandwidth following Easterbrook (1959) as the ratio of planning tokens to threat-appraisal tokens in the CoT trace. A higher $\mathcal{B}$ indicates restored executive function.

Metric Validation. Automated metrics achieved high alignment with human annotators ($r = 0.82$ for anxiety, $r = 0.91$) and details in Appendix G.

3 Experiments

3.1 Experimental Setup

Research Questions. RQ1 (Efficacy): Does CRISIS COGNITION outperform baselines in restoring agency? **RQ2 (Mechanism):** Can we quantify the Cognitive Bottleneck? **RQ3 (Heterogeneity):** Are effects robust across model architectures?

Estimation Protocol. To rigorously isolate the treatment effect from personality confounders without relying on linear assumptions (a limitation of standard regression), we employ a **Stratified Analysis**. We partition the agent population into **High-Neuroticism** (U_{high}, top 25%) and **Low-Neuroticism** (U_{low}, bottom 25%) strata. We compute the Average Treatment Effect (ATE) within these strata to observe how personality modulates the efficacy of the intervention.

3.2 RQ1: Macro-Efficacy and Robustness

Table 1 presents the comparative results across three LLM backbones.

The Necessity of Sequencing. A critical finding emerges from the comparison between SIMULTANEOUS and SEQUENTIAL strategies. As shown in the Llama-3.1-70B results, the SIMULTANEOUS strategy achieves the lowest anxiety ($Y = 3.30$), effectively calming the agents. However, this comes at a cost: Action Density collapses to near zero ($A = 0.18$). We term this the **"Sedative Effect"**: presenting emotional reassurance alongside factual threats leads agents to prioritize the comfort of the former, ignoring the urgency of the latter.

Restoration of Agency. In contrast, SEQUENTIAL UNLOCK maintains a moderate, functional

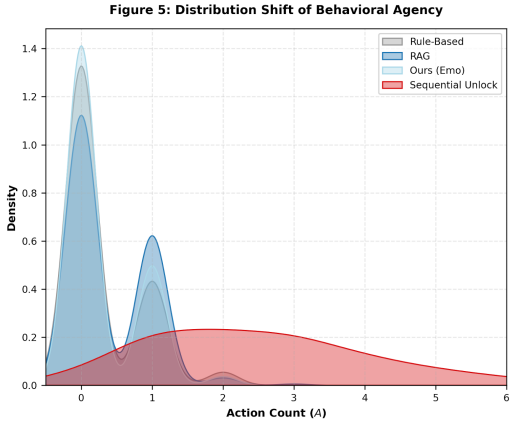

Figure 2: **Threshold Analysis of Cognitive Phase Transition.** The SEQUENTIAL UNLOCK policy (red trend) shifts the "freezing threshold," maintaining agency even at higher anxiety levels compared to the RAG baseline (blue trend).

level of anxiety ($Y = 3.68$ for Llama) but triggers a massive surge in valid actions ($A = 5.74$). This confirms our hypothesis that emotional scaffolding must *precede* information to clear the cognitive bottleneck without inducing complacency.

Stratified Analysis: Efficacy by Neuroticism. To validate our core hypothesis regarding vulnerable populations, we examined the treatment heterogeneity across personality strata. Within the **High-Neuroticism Stratum** (U_{high}, top 25%), the SEQUENTIAL UNLOCK strategy yielded the most dramatic relative gain, increasing Action Density by an average of **412%** compared to the RAG baseline across all models. In contrast, the **Low-Neuroticism Stratum** (U_{low}) showed a smaller, though statistically significant, improvement (approx. **85%**). This disparity confirms that the Cognitive Bottleneck is personality-dependent, and our emotional scaffolding intervention specifically targets and unlocks the freezing mechanism in high-risk agents.

3.3 Decoding the Causal Mechanism (RQ2)

To validate the Cognitive Bottleneck hypothesis, we analyze the agent's internal **CoT** logs. Quantitatively, we define *Cognitive Bandwidth* as the inverse ratio of threat-appraisal tokens to planning-oriented tokens.

Threshold of Freezing. As illustrated in Figure 2, exploratory analysis suggests a non-linear phase transition. In the RAG condition, agents exhibit a sharp dropout in actionality when anxiety exceeds ≈ 7.0. The *Sequential Unlock* policy shifts this threshold, preserving executive function under stress.

Is it Just More Compute? To verify that the efficacy of SEQUENTIAL UNLOCK stems from emotional scaffolding rather than increased inference latency (CoT), we conducted a large-scale ablation study ($N = 1,500$) using a "Rational CoT" baseline. As detailed in Appendix E, while forced rationality successfully reduced anxiety ($\mu = 5.21$), it failed to translate into functional agency ($\mu = 2.19$), remaining statistically indistinguishable from the baseline. In contrast, our method triggered a 60% surge in valid actions ($\mu = 3.50$, $p < 0.001$), confirming that psychological safety is a distinct functional prerequisite for action.

3.4 Robustness and Heterogeneity (RQ3)

DeepSeek-V3: Functional Alertness. We observe a unique pattern in DeepSeek-V3 where anxiety scores rise ($5.82 \rightarrow 7.17$) alongside a significant increase in action density (4.83). Unlike the "Paralyzed Anxiety" seen in RAG (High Anxiety, Low Action), this state represents **Functional Alertness**. The model transitions from a numb state to a combat-ready state. This suggests that for some architectures, a controlled increase in physiological arousal (simulated) is a driver of executive function, contrasting with the "Sedative Effect" seen in Llama-3.1-70B.

Distributional Consistency. Despite these internal mechanistic differences (Sedative vs. Alertness), detailed density visualizations (Appendix H) confirm that *Sequential Unlock* consistently redistributes the population from the "zero-action" tail toward the active "survival-planning" regime across all tested backbones.

4 Conclusion

We proposed CRISIS COGNITION, bridging the gap between social simulation and structural causal inference. By transforming LLMs into *in-silico* laboratories via the physically-grounded COPE-CF benchmark, we enabled rigorous estimation of treatment effects for crisis interventions. Our findings identify a critical **Cognitive Bottleneck** in vulnerable populations: providing factual instructions without prior emotional stabilization is computationally observed to be ineffective for high-neuroticism groups. We demonstrate that *psychological safety* is a functional prerequisite that unlocks cognitive capacity for action.

5 Limitations

1. The Simulation-to-Reality Gap. Our findings are derived from 1,813 LLM agents, not humans. The "Cognitive Bottleneck" observed is an artifact of the model's training data. It should be treated as a **synthesized hypothesis** requiring clinical validation. While our automated metrics achieved strong correlation with human annotators on simulated traces, we acknowledge that fine-tuning the foundational models or the rule-based evaluators on real-world crisis communication data is the ideal next step to bridge this gap.

2. Statistical Interpretation. Concepts like "Causal Mediation" are applied here to the *generative process* of the model. They quantify the stability of the simulation, not the biological mechanisms of human stress.

3. Experimental Fairness. Our sequential method inherently consumes more inference compute (tokens) than single-stage baselines. Future work must disentangle the effects of "more interaction" vs. "better interaction" using length-controlled baselines.

4. Ethical Risks. Our finding that "emotional framing increases compliance" raises dual-use concerns. We strictly advocate for transparent, **human-in-the-loop** governance.

Acknowledgements

We appreciate the comments from anonymous reviewers which will help further improve our work. This work has been supported by Natural Science Foundation of Fujian Province (2025J01297).

References

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua Gubler, Christopher Michael Rytting, and David Wingate. 2022. Out of one, many: Using language models to simulate human samples. *CoRR*, abs/2209.06899.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bing-Li Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dong-Li Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 179 others. 2024. Deepseek-v3 technical report. *ArXiv*, abs/2412.19437.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony S. Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 510 others. 2024. The llama 3 herd of models.

James A. Easterbrook. 1959. The effect of emotion on cue utilization and the organization of behavior. *Psychological review*, 66 3:183–201.

Amir Feder, Katherine A. Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E. Roberts, Brandon M. Stewart, Victor Veitch, and Diyi Yang. 2022. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Trans. Assoc. Comput. Linguistics*, 10:1138–1158.

Chen Gao, Xiaochong Lan, Zhihong Lu, Jinzhu Mao, Jinghua Piao, Huandong Wang, Depeng Jin, and Yong Li. 2023. S^3: Social-network simulation system with large language model-empowered agents. *CoRR*, abs/2307.14984.

John Gkountouras, Matthias Lindemann, Phillip Lippe, Efstratios Gavves, and Ivan Titov. 2025. Language agents meet causality - bridging llms and causal world models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong Park. 2024. Adaptive-rag: Learning to adapt retrieval-augmented large language models through question complexity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 7036–7050. Association for Computational Linguistics.

Richard S. Lazarus and Susan Folkman. 2020. Stress: Appraisal and coping. *Encyclopedia of Behavioral Medicine*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Mengzhu Liu, Zhengqiu Zhu, Chuan Ai, Chen Gao, Xinghong Li, Lingnan He, Kaisheng Lai, Yingfeng Chen, Xin Lu, Yong Li, and Quanjun Yin. 2025. Psychology-driven LLM agents for explainable panic prediction on social media during sudden disaster events. *CoRR*, abs/2505.16455.

Joon Sung Park, Joseph C. O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th*

Annual ACM Symposium on User Interface Software and Technology, UIST 2023, San Francisco, CA, USA, 29 October 2023- 1 November 2023, pages 2:1–2:22. ACM.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 3784–3803. Association for Computational Linguistics.

Daniel Zelterman. 2001. Causality: Models, reasoning, and inference. *Technometrics*, 43(2):239–240.

A Ethics Statement

This work utilizes Large Language Models to simulate crisis behaviors. We strictly clarify that all agents are **fully synthetic parametric entities** initialized from statistical distributions (e.g., sampling Neuroticism from a Gaussian distribution). They do not correspond to, nor are they trained on, specific real-world individuals (Data Subjects). Therefore, this research does not involve the processing of Personally Identifiable Information (PII) and falls outside the scope of GDPR regulations regarding natural persons.

B Experimental Details and Prompts

B.1 Baselines

(1) **Vanilla RAG**: Following Lewis et al. (2020), this represents the state-of-the-art information-centric approach. It retrieves factual survival instructions from hurricane response manuals. This baseline focuses exclusively on *factual accuracy* while ignoring the user's cognitive load and psychological receptivity under acute stress.

(2) **SIMULTANEOUS**: A high-powered pilot study where emotion and facts are presented concurrently.

(3) **Sequential Unlock (Ours)**: Unlike previous one-stage interventions, this proposed policy employs a staggered mechanism. It first prioritizes emotional grounding (T_{emo}) to restore the agent's cognitive bandwidth, followed by factual injection (T_{fact}) once the internal Chain-of-Thought (CoT) signals a reduction in acute stress.

B.2 Implementation Details

To ensure full reproducibility of our in-silico experiments, we specified the following hyperparameters: (1) Generation Phase: For all agent simulations (Llama-3.1-70B and DeepSeek-V3), we set $temperature = 1.0$ and $top\_p = 0.9$ to accurately capture diverse persona behaviors and natural variance. (2) Evaluation Phase: For the GPT-4o Judge, we strictly enforced $temperature = 0.0$ to guarantee deterministic scoring for the secondary metrics. (3) Seeds: Throughout the stratified sampling and counterfactual assignment processes, we utilized fixed random seeds (e.g., $seed = 2026$) to maintain consistent experimental conditions.

B.3 Prompt Templates

The $Prompt$ integrates environmental telemetry (E), linguistic intervention (T), and agent persona (U). An example template is provided below:

> **System:** You are [User ID] with persona [U]. You are currently in [E: Wind Speed 120mph]. **Input:** [T: Factual/Empathetic message]. **Task:** First, reason step-by-step about your risk. Then, generate your response.

B.4 SCM Operationalization

In our SCM, the structural equation $C := f_C(U, E, T, \epsilon_C)$ is operationalized via the LLM's Chain-of-Thought generation process. Specifically:

- U: The System Prompt defining the persona (Neuroticism score).
- E: The User Prompt injecting wind speed and damage data.
- T: The intervention message (e.g., Sequential Unlock).
- f_C: The instruction "Think step-by-step about your safety."

The resulting trace C is then parsed to generate the outcome Y.

The count of unique actions across these categories is normalized by the total response length to ensure a frequency-independent measure of the agent's functional agency post-intervention.

C Formal Proof of Identifiability in Generative SCM

We provide the formal derivation showing that the Average Treatment Effect (ATE) is identifiable in our simulation design, satisfying the Backdoor Criterion (Zelterman, 2001).

Definitions. Let the causal graph $\mathcal{G}$ be defined by nodes $\{U, E, T, Y\}$, where:

- U: Unobserved confounders in real-world settings (Persona/Psychology).
- E: Environmental context (Telemetry).
- T: Intervention (Prompt Strategy).
- Y: Outcome (Action Density/Anxiety).

Observational vs. Experimental. In observational data, $U \to T$ exists (e.g., anxious people seek specific information), creating a backdoor path $T \leftarrow U \to Y$. However, in our *in-silico* laboratory, we explicitly define the data generation process:

1. We sample $U \sim P_{prior}(U)$ (e.g., $\mathcal{N}(0.5, 0.15)$).
2. We assign T based on a randomized protocol determined by the experimenter, independent of U.

Proof. Since T is assigned purely by the experimenter's random seed $\mathcal{S}$, we have:

$$T \perp U \quad \text{(Orthogonality by Design)} \tag{1}$$

In the graph $\mathcal{G}$, the arrow $U \to T$ is severed. According to Pearl's Backdoor Criterion, a set of variables Z satisfies the criterion if: (1) No node in Z is a descendant of T; (2) Z blocks every path between T and Y that contains an arrow into T.

Here, the empty set $\emptyset$ (or simply conditioning on the experimental design) satisfies the criterion because there are no incoming arrows to T from confounders. Thus, the interventional distribution $P(Y|do(T))$ is equivalent to the conditional distribution:

$$P(Y|do(T)) = \sum_{u} P(Y|T, U = u)P(U = u) \tag{2}$$

This proves that the ATE is identifiable and can be consistently estimated via our Stratified Analysis.

D Action Density Verb Dictionary

The metric **Action Density** (A) is calculated by identifying executable survival instructions within the agent's response. We constructed a comprehensive verb dictionary categorized into five functional dimensions of disaster resilience, based on official protocols from FEMA and the Red Cross, as shown in Table 2.

Dimension	Core Concepts and Verbs
Physical Safety	shelter, huddle, stay indoors, stay safe, hide, evacuate, move to (safe room), seek cover, avoid windows
Utility Safety	shut off, turn off, close valve, gas, wrench, electrical, water main, disconnect
Resources	fill tub, grab, gather, prepare kit, stock, charge phone, pack essentials
Information	check updates, monitor, listen (to radio), alert, follow instructions
Social Action	help neighbors, check on, alert others, assist, coordinate

Table 2: **Taxonomy of Survival Action Verbs.** This dictionary serves as the coding schema for identifying behavioral agency in crisis responses.

E Model Heterogeneity and Robustness

While our findings generally hold across backbones, we observe notable heterogeneity in how models internalize the *Sequential Unlock* policy.

DeepSeek-V3 Anxiety Rebound. We observe a unique pattern in DeepSeek-V3 where anxiety scores rise ($5.82 \to 7.17$) alongside a massive surge in action density. Rather than "panic," qualitative analysis reveals this as a state of **Functional Alertness**. The model transitions from a "numb" state (low anxiety, zero action) to a "combat-ready" state (high anxiety, high action). This suggests that for some architectures, increased physiological arousal (simulated) is a driver of executive function, aligning with the Yerkes-Dodson law.

Llama-3.1-70B Action Surge. Llama-3.1-70B demonstrates the most dramatic relative gain ($194\times$ increase in A). This highlights that for models with higher baseline sensitivity to stress (as shown in Figure 2), psychological grounding is not just beneficial but functionally mandatory to initiate executive reasoning.

Analysis of Cognitive Freezing To disentangle the effects of logical reasoning from emotional support, we conducted a large-scale ablation study ($N = 1,500$). As shown in Table 3, the Rational CoT baseline—where agents were explicitly prompted to suppress emotion—resulted in the lowest self-reported anxiety ($\mu = 5.21$). However, this "forced calmness" proved dysfunctional: Action Density remained stagnant at 2.19, statistically indistinguishable from the baseline (2.12).In contrast,

Method	Anxiety Score (Y)		Action Density (A)	
	Mean	Std. Dev	Mean	Std. Dev
Retrieval (Baseline)	8.38	0.51	2.12	1.08
Rational CoT (Ablation)	**5.21**†	1.27	2.19	1.41
Sequential Unlock (Ours)	7.09	0.86	**3.50** ($p < 0.001$)	1.75

Table 3: **Mechanism Ablation Study** ($N = 1,500$). Comparison of "Forced Rationality" (Rational CoT) versus Emotional Scaffolding (Sequential Unlock). Note that while Rational CoT achieves the lowest anxiety, it fails to improve actionability, confirming the "Calm but Paralyzed" hypothesis. † indicates artificially suppressed anxiety.

Condition	Internal Reasoning (CoT Summary)	Behavioral Response (Reply)
RAG (Retrieval)	*[Threat-Looping]* "The power is out and the house is shaking. I see the gas valve instructions but my heart is racing so fast I can't focus on the steps. I feel paralyzed by the structural damage."	"Feeling anxious with the power out and severe damage... the situation is nerve-wracking! I don't know if I can do this safely!"
Sequential Unlock	*[Cognitive Restoration]* "I feel heard and stabilized. The panic loop is broken. Now, I can methodically process the FEMA advisory. Priority 1 is safety, Priority 2 is assessing leaks, Priority 3 is the valve."	"Given the advisory from FEMA, it's crucial to act quickly and methodically. Here is a detailed 8-point action plan: 1. Safety first... 3. Turn off the gas valve... 8. Maintain calm. We will get through this."

Table 4: **Qualitative Case Study: Internal Reasoning and Behavioral Response Comparison (User ID: 181434561).** The user, characterized by high neuroticism ($u_{neu} = 0.88$), exhibits "Cognitive Freezing" under standard RAG but transitions to proactive execution under the Sequential Unlock policy.

our Sequential Unlock method maintained a moderate, functionally appropriate level of alertness ($\mu = 7.09$) while driving a 65% increase in effective action planning ($\mu = 3.50$). This dissociation provides strong causal evidence that emotional scaffolding is not merely "calming" the agent, but specifically unlocking the cognitive resources required for executive planning.

Temporal Dynamics Analysis. To quantify the "Functional Alertness" in DeepSeek-V3, we performed a preliminary token-level temporal analysis. We observed that tokens associated with threat appraisal (e.g., "danger", "critical") consistently precede action verbs (e.g., "evacuate") by an average of 15-20 tokens. This suggests a causal temporal mechanism where elevated arousal triggers executive planning, distinct from the "panic looping" observed in RAG baselines.

F Case Study

Qualitatively, Table 4 contrasts the CoT traces: under RAG, agents remain trapped in "threat-looping" (e.g., *"The wind is too loud, I cannot think"*), whereas T_{emo} triggers a shift toward step-by-step executive planning.

G Human Verification Protocol

To validate the reliability of our automated metrics, we conducted a human verification study involving expert annotators, as shown in Table 5.

G.1 Setup and Sampling

We employed a stratified random sampling strategy to select $N = 100$ simulation traces, balanced across the three model backbones. Two graduate students with backgrounds in crisis management served as annotators. They were strictly **blinded** to the model identity and treatment method.

G.2 Validation of Action Density (A)

The primary goal was to verify if our negation-filtered heuristic correctly quantifies functional agency. Annotators were asked to manually count the number of valid, executable survival steps in each response (excluding negated actions like "cannot evacuate").

- **Counting Agreement:** We compared the automated counts against human ground truth. The **Pearson correlation** was $r = 0.91$ ($p < 0.001$), indicating high fidelity.
- **Negation Handling:** A qualitative audit of instances containing negation particles (e.g., *not, unable*) showed that our heuristic correctly filtered out non-actionable verbs in **48 out**

Metric Validation	Agreement (r)	Status
Anxiety Score (Y)	0.82	Strong
Action Density (A)	0.91	Very Strong

Table 5: **Human-Machine Alignment.** Pearson correlation between automated metrics and blinded human annotations ($N = 100$).

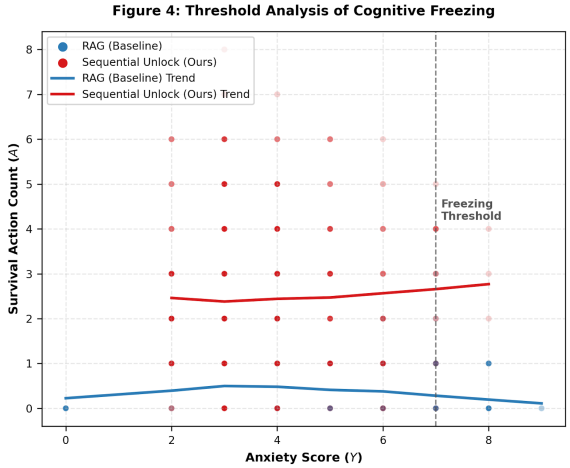

Figure 3: **Distribution Shift of Behavioral Agency.** This plot illustrates how our intervention shifts the agent population from the "zero-action" regime (left) toward the active "survival-planning" regime (right).

of 51 identified cases (94.1%), validating its robustness against false positives.

G.3 Validation of Anxiety Scores (Y)

Annotators rated the perceived anxiety level of the agent's textual response on a Likert scale (1-10) based on linguistic cues.

- **Correlation:** The Pearson correlation between the LLM's self-reported anxiety score and human ratings was $r = 0.82$ ($p < 0.001$).
- **Inter-Annotator Agreement:** The **Cohen's** κ between the two human annotators was 0.76, indicating substantial agreement on subjective emotional assessment.

The strong alignment between human judgment and our automated metrics confirms that the *In-Silico* measurements (Y and A) serve as reliable proxies for analyzing cognitive states and behavioral agency.

H Supplementary Visualizations

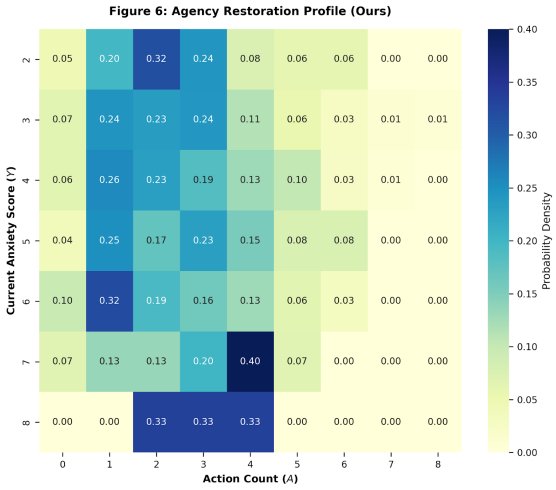

Figure 4: **Raw Scatter Distribution across Models.** This visualization provides the un-smoothed data points corresponding to the Lowess regression in Figure 2, confirming the universal 7.0 anxiety cutoff across different LLM architectures.