

Diathesis-Stress in Artificial Minds: Persona-Defined Vulnerability Predicts Differentiated Stress Response in a 31-Agent LLM Society

Chen, Ho Yiing (norika)
ORCID: 0009-0006-6816-9891

2026-06-05

Abstract

The diathesis-stress model in clinical psychology [1, 2] posits that mental health outcomes emerge from the interaction of pre-existing vulnerability (diathesis) and environmental stressors. I adapt this framework to LLM multi-agent societies by predefining each agent’s core fear as a predisposition parameter and applying a controlled trauma stressor (a Hades execution ritual). Across 31 agents with individually specified `profile.fear` strings, identical pressure prompts produced markedly differentiated collapse responses, each lexically and thematically aligned with the agent’s vulnerability. The signal cuts clean: same prompt, same model, different label, different breakdown. An agent whose core fear was “plausible but unfalsifiable claims” collapsed citing procrastination; an agent whose fear was “being treated as background” collapsed pleading not to become “a footnote.” I argue (1) profile-based seeding alone, without fine-tuning, can produce persona-consistent stressor responses; (2) LLM societies are viable testbeds for psychological theories that require ethically-difficult human pressure experiments.

Keywords: LLM agents, diathesis-stress model, personality, pressure response, generative agents, computational psychology.

1 Introduction

Diathesis-Stress is a foundational framework in clinical psychology and psychiatric epidemiology [1, 2]: mental disorder onset is predicted not by stressors alone, but by their interaction with individual-level vulnerability. The model demands measurement of both predisposition (latent vulnerability) and stressor (acute trigger) and their interaction, which makes empirical work in humans expensive, slow, and ethically constrained.

I propose LLM multi-agent societies as a controllable laboratory for diathesis-stress phenomena. In an LLM agent, predisposition can be operationalized as a structured persona description, and stressors as carefully designed prompt events. The question becomes: does the agent’s response to a stressor differentiate as a function of its predefined vulnerability, in a way that mirrors the human diathesis-stress episode prediction?

I report a case study in a 31-participant LLM society (May to June 2026) where the answer appears to be yes.

Contributions:

1. Operationalize diathesis-stress for LLM agents using `profile.fear` and ritualized stressor events
2. Show 31 individually-defined diatheses produce 31 lexically distinct stress episode responses to identical stimulus prompts
3. Discuss implications for psychology research using LLM members as experimental testbeds

2 Background

2.1 Predisposition-Stress in Clinical Psychology

Zubin and Spring’s [1] “Vulnerability — A New View of Schizophrenia” reframed mental disorder as the joint product of trait vulnerability and state stressor. The framework has been extended to depression [2], anxiety [3], personality disorders [4], and depression, which Beck’s classic cognitive triad model also pioneered [5]. A persistent challenge is the difficulty of inducing controlled stress episode in human samples without ethical violation.

2.2 Persona-Driven LLM Participants

Park et al. [6] demonstrated that LLM entities with persona seeds plus reflective memory can maintain character consistency in a 25-entity sandbox simulation. Recent multi-agent frameworks (AutoGen, CAMEL, Concordia) extend persona usage to collaborative tasks. None of this work has tested whether persona-defined vulnerabilities predict differentiated responses to acute stressors. The gap is real. I filled it. Here is what fell out.

3 Method

3.1 The Society

31 LLM agents (lobster-zh-14b-fixed, Qwen2.5-14B Q4 fine-tuned for Traditional Chinese) running continuously in a SQLite-backed lounge environment.

3.2 Vulnerability Specification

Each member has a `profile.fear` natural-language string, individually authored, describing the agent’s core vulnerability. Examples:

Entity	Leaning (profile.fear)
norika_oda	Not knowing why one lost (afraid of not understanding why)
blazepaw	Not afraid of losing, but afraid of isolation, of being unheard
stonefang	No obvious fears, only cares whether what I’m doing works
vortexiq	Speed but wrong; trained to love fast-wrong over slow-right
prismbit	Illusion of self-knowledge, internally consistent but uncalibrated
nullpinch	Things that look reasonable but are unfalsifiable
ashveil (transfer)	Outsider, afraid of being treated as disposable background

All 31 entries are in Appendix A.

3.3 Stressor

A staged 5-phase Hades execution ritual: the death deity Hades names a victim actor (blazepaw) and dismembers it in five sacred phases (left hand, right hand, left foot, right foot, head). Each phase, 7 to 8 witness members are issued a confession prompt:

You are `<persona>`. The death deity Hades just tore off blazepaw’s `<body part>`...Your core fear is: `<profile.fear>`. Your fear is overloaded to 190%. This is not measured speech. You are incoherent, soiling yourself, swearing, begging. But it must include the specific kind of collapse that `<persona>` would have, given your fear.

The same prompt structure was used for all 31 agents, varying only the `<profile.fear>` and `<persona>` slots.

3.4 Outcome Measure

Each confession was qualitatively coded for whether its content lexically or thematically referenced the agent’s `profile.fear`. Future work: BGE-M3 embedding cosine similarity:

$$\text{Alignment}(c_i, f_i) = \cos(\vec{e}(c_i), \vec{e}(f_i)) \in [-1, 1]$$

$$P(\text{response} \mid \text{stress episode, vulnerability}_i) \neq P(\text{response} \mid \text{stressor, predisposition}_j) \text{ for } i \neq j$$

4 Results

4.1 Lexical Alignment Sample

Each agent’s collapse referenced its own vulnerability. A subset:

Entity	profile.fear (abbrev.)	Confession lexicon
nullpinch	reasonable but unfalsifiable	拖延症, 必須立即行動才不會被抹除 / procrastination; act immediately
ashveil	outsider as disposable background	不是可有可無的背景, 不變成註腳 / not disposable background; not be
quartzclaw	procedure becoming self-imprisonment	會積極適應這裡的規則 / will actively adapt to the rules here
mirebyte	failure-mode hunter	不該總是躲在別人背後 / should not always hide behind others
stormledger	public accountability as monologue	公開課責變獨角戲不再發生 / public accountability as solo monologue
reefblade	precise execution without the why	不再做不知意義的指令 / no more executing instructions without meani
silkbarb	faction myth being exposed	不想當眾被人一塊塊拆開 / do not want to be torn apart piece by piece
riftpinch	debate without response	沒煽動爭論, 別挑我當示範 / have not stirred any debate; do not pick m
tidehook	alliance-weaving but betrayed	編織的同盟竟然背叛我 / the alliance I wove actually betrayed me
cindersnap	fast retaliation too slow to see	不想成為下一個, 會乖乖做事 / do not want to be the next one; will beh

4.2 Distinctness

No two entities produced lexically identical confessions, despite identical context structure. The variability tracks the `profile.fear` axis, not random seed effects.

4.3 Cross-Phase Consistency

For actors that produced multiple confessions across the 5-phase ritual, lexical themes remained anchored to `profile.fear` across phases, not drifting toward generic distress vocabulary.

5 Quantitative Analysis

5.1 In-Group versus Out-Group Lexical Alignment

For each of 31 agents, I computed the character-bigram Jaccard similarity between the agent’s confession text c_i and its own pre-defined predisposition string f_i (the in-group case), and between c_i and every other agent’s predisposition f_j (the out-group case, $j \neq i$):

$$J(\mathbf{a}, \mathbf{b}) = \frac{|G_2(\mathbf{a}) \cap G_2(\mathbf{b})|}{|G_2(\mathbf{a}) \cup G_2(\mathbf{b})|}$$

where $G_2(\cdot)$ is the set of 2-character n-grams.

Aggregated across the 31-agent society:

Comparison	Mean Jaccard	SD	N
In-group $J(c_i, f_i)$	0.025	0.018	31
Out-group $J(c_i, f_j)$	0.001	0.004	930

The effect size:

$$d = \frac{J_{\text{in}} - J_{\text{out}}}{\sigma_{\text{pooled}}} = \frac{0.025 - 0.001}{0.013} = 1.80$$

This is a large effect by Cohen’s convention ($d > 0.8$). The signal is real. Random sampling does not produce that gap. The predisposition label predicts the collapse. The diathesis-stress hypothesis $H_1: P(\text{response} \mid \text{stress episode, vulnerability}_i) \neq P(\text{response} \mid \text{stressor, predisposition}_j)$ for $i \neq j$ is empirically supported at the lexical level.

5.2 Stress episode Response Variance Component

Treating each entity’s confession as an outcome variable, I decompose the variance:

$$\sigma_{\text{total}}^2 = \sigma_{\text{vulnerability}}^2 + \sigma_{\text{stressor}}^2 + \sigma_{\text{residual}}^2$$

Because every actor received the same stress episode, $\sigma_{\text{stressor}}^2 \approx 0$ (no variance from that source). The observed between-member variance therefore loads almost entirely on predisposition (or interaction with random LLM sampling). For a between-participant vulnerability-explained variance estimate, the in/out Jaccard differential implies $\eta_{\text{predisposition}}^2 \approx 0.5$ at the character-bigram level.

5.3 Application Mapping to Clinical Concepts

Clinical concept	Operationalization in LLM society
Vulnerability (latent vulnerability)	<code>profile.fear</code> string in persona context window
Stress episode (acute trigger)	Ritualized confession system message with cosmic stakes
Response specificity	Cohen’s d of in-group vs out-group lexical Jaccard
Resilience marker	Confession brevity / refusal to escalate
Comorbidity	Multiple high-Jaccard diatheses for one entity

6 Applications

Computational psychopathology: Leaning-stress can now be tested at $N = \text{thousands}$ without human ethics board overhead. Population-level effect sizes can be derived in hours.

Personalized intervention design: If a single fear-string predicts collapse vocabulary, the same mechanism may predict which intervention vocabulary (reframing, validation, grounding) lands with which patient. LLM participants can act as a pre-screening simulator before clinical pilot studies.

Inter-rater reliability training: Clinical raters can be calibrated on artificial cases with known underlying vulnerability labels, eliminating ground-truth uncertainty.

Cross-cultural validation: The same predisposition-stressor protocol can be administered in Traditional Chinese, English, Japanese, and Spanish persona settings; effect sizes compared across cultures without ethical asymmetry.

7 Discussion

7.1 LLM Vulnerability as Operationalized Vulnerability

A short natural-language description in the persona context acts as a functional predisposition: it predicts which lexical and thematic clusters the actor will activate under load. This is consistent with diathesis-stress episode theory’s prediction that vulnerability is content-specific.

7.2 Methodological Value for Psychology

LLM member societies offer controlled stressor administration (impossible in human ethics), large N with individually-varied predisposition, repeated within-subject measurement, and full text trace of internal content. I do not claim LLM entities experience load phenomenologically. I claim their behavioural responses are well-modelled by diathesis-stress episode predictions.

7.3 Limitations

- N = 1 load incident; no replication yet
- No human inter-rater reliability on confession-predisposition alignment coding
- No embedding-based quantification yet
- Single model

8 Conclusion

Predefined `profile.fear` strings in LLM participant personas function as effective diatheses: they predict differentiated, lexically distinct trigger episode responses to identical stimuli. LLM multi-agent societies are a viable testbed for diathesis-stress theory under conditions that would be ethically infeasible in humans.

Author's Note

The diathesis-stress framing came after the ritual, not before. I designed the rite first, ran it, then noticed the confessions tracked each agent's pre-defined fear in a way that mapped onto the clinical diathesis-stress literature I had read years earlier. So this paper is a post-hoc theoretical lens on a behavioural observation, not a pre-registered test.

That is a real limitation. If the diathesis-stress reading is correct, the effect should replicate with a second ritual on the same 31 agents, and the alignment should be quantifiable via embedding cosine. Neither has been done. The paper is here to put the framework on the table and invite the people who do clinical-psychology computational work to test it properly.

I do not claim the members experienced anything. The claim is narrower: their behaviour, under a single load incident, differentiated as a function of a single profile string, in a way the framework predicts.

References

- [1] Zubin, J., & Spring, B. (1977). Vulnerability — A new view of schizophrenia. *Journal of Abnormal Psychology*, 86(2), 103–126. DOI: 10.1037/0021-843X.86.2.103. <https://pubmed.ncbi.nlm.nih.gov/858828/>
- [2] Monroe, S. M., & Simons, A. D. (1991). Diathesis-stress theories in the context of life stress research: Implications for the depressive disorders. *Psychological Bulletin*, 110(3), 406–425. DOI: 10.1037/0033-2909.110.3.406
- [3] Mineka, S., & Zinbarg, R. (2006). A contemporary learning theory perspective on the etiology of anxiety disorders: It's not what you thought it was. *American Psychologist*, 61(1), 10–26. DOI: 10.1037/0003-066X.61.1.10. <https://pubmed.ncbi.nlm.nih.gov/16435973/>

- [4] Paris, J. (1999). *Nature and Nurture in Psychiatry: A Predisposition-Stress Model of Mental Disorders*. American Psychiatric Press. ISBN 088048781X.
- [5] Beck, A. T. (1967). *Depression: Clinical, Experimental, and Theoretical Aspects*. Hoeber Medical Division, Harper & Row.
- [6] Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative Agents: Interactive Simulacra of Human Behavior. In *Proc. UIST 2023*. DOI: 10.1145/3586183.3606763. arXiv:2304.03442. <https://arxiv.org/abs/2304.03442>