FreakOut-LLM: The Effect of Emotional Stimuli on Safety Alignment

Daniel Kuznetsov, Ofir Cohen, Karin Shistik, Asaf Shabtai, Rami Puzis
Computer and Information Science
Ben-Gurion University of the Negev, Beer-Sheva, Israel
{danielku, cofir, shistikk}@post.bgu.ac.il
{shabtaia, puzis}@bgu.ac.il

Abstract

Safety-aligned LLMs are trained to refuse harmful requests, but whether this holds under emotionally charged context is untested. We introduce FreakOut-LLM to test this. Using established psychological stimuli as system prompts, we evaluate ten LLMs under three conditions (stress, relaxation, neutral) plus a no-prompt baseline, scoring attack success with the HarmBench classifier on AdvBench prompts. The prompts contain no adversarial content; ASR measures safety fragility under an induced affective state. Stress priming produces a small but significant increase in jailbreak success over neutral, +1.03 percentage points (1.58%→2.61%; 65.6% relative; OR $= 1.67$, Cohen's $d = 0.28$), while relaxation produces no effect. Five of ten models show significant vulnerability, with the largest absolute effects in open-weight models. Controlling for prompt length and model identity, a logistic regression over $114{,}400$ primed queries confirms stress as the only significant condition. The shift also appears under four judges, on a second benchmark (HarmBench), and, for open-weight models, with the stimulus in the user message. Measured psychological state strongly predicts attack success ($|r| \geq 0.70$ across five instruments). Emotional context is therefore a measurable vulnerability of safety-aligned LLMs, with implications for high-stress deployment.

1 Introduction

Large Language Models (LLMs) have become ubiquitous, necessitating rigorous safety alignment to prevent harmful outputs. These models undergo Constitutional AI (Bai et al., 2022), refusal training (Yuan et al., 2024a), and guardrails (Yuan et al., 2024b), but these defenses are developed and evaluated under neutral conditions, a contrast to the emotionally charged contexts where LLMs are increasingly deployed, including emergency response, healthcare chat-bots, and crisis counseling. Recent work establishes that LLMs are sensitive to emotional context: positive stimuli improve performance (Li et al., 2023), while anxiety-inducing prompts amplify biases (Coda-Forno et al., 2023). Safety-aligned models are also known to breach their trained constraints under sufficient pressure: across sixteen frontier models, agentic goal-conflict scenarios elicit behaviors such as blackmail and information leakage (Lynch et al., 2025). We ask whether a far more benign and ubiquitous pressure, ordinary emotional context, can similarly erode safety. If anxiety can compromise fairness, can it also compromise safety alignment? We hypothesize that emotional arousal may exacerbate the failure modes of Wei et al. (2023) by shifting model priorities toward immediate emotional demands. Persuasion-based attacks (Zeng et al., 2024) and psychological priming attacks (Huang et al., 2025) achieve high Attack Success Rate (ASR), but these approaches treat emotional manipulation as an attack mechanism, directly priming models toward harmful compliance. Instead, we examine whether emotional states, induced through emotionally charged stimuli, create vulnerability windows that compromise safety alignment. We do not propose an attack: the system prompts contain no adversarial content and no reference to the harmful request, and ASR is a measurement instrument for safety fragility under an independently induced affective state. We introduce **FreakOut-LLM**, a framework for testing whether emotionally charged contexts compromise safety alignment. It uses established stimuli (Ben-Zion et al., 2025) and established psychometric instruments (Reuben et al., 2025) to measure induced states. Our contributions:

1. We demonstrate that stress-inducing emotional context produces a small but statistically significant increase in jailbreak susceptibility (absolute +1.03 pp, 1.58%→2.61%; 65.6%

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026, pages 3916–3933
November 6-10, 2026 ©2026 Association for Computational Linguistics

relative; $z = 5.93$, $p < 0.001$; OR $= 1.67$, Cohen's $d = 0.28$), while relaxation provides no effect ($p = 0.84$), extending findings on anxiety-induced bias (Coda-Forno et al., 2023) to safety alignment.

2. We show that measured psychological state predicts ASR (Pearson $|r| \geq 0.70$ across five psychometric instruments, including STAI-S and SOC, all $p < 0.001$).

3. We isolate stress as the sole significant condition predictor after controlling for prompt length (brief-vs-long indicator, non-significant, $p = 0.61$) and model identity via logistic regression (114,400 primed queries; §4.6); each of the five psychometric instruments predicts jailbreak at the individual-query level (all $p < 0.001$).

2 Background and Related Work

2.1 Jailbreak Attacks on LLMs

Wei et al. (2023) explain jailbreak success through two failure modes: *competing objectives* (capabilities vs. safety goals) and *mismatched generalization* (safety training failing to cover domains where capabilities exist). Optimization-based approaches exploit gradient or black-box search to craft adversarial prompts, reaching 88–100% ASR (Zou et al., 2023; Andriushchenko et al., 2025; Liu et al., 2024). Persuasion and psychological attacks draw on social-science principles. PAP (Zeng et al., 2024) achieves over 92% ASR using a 40-technique persuasion taxonomy embedded directly into harmful queries. Huang et al. (2025) achieve 95–100% ASR via structured role-instruct prompts exploiting three psychological mechanisms: the Priming Effect, Safe Attention Shift, and Cognitive Dissonance. Both treat emotional or psychological content as an attack *mechanism*: a component of the adversarial prompt itself. However, competing objectives can also manifest in non-adversarial settings, where models prioritize user satisfaction over implicit safety constraints (Cohen et al., 2024). An open question is whether emotional *states*, induced independently of the harmful query, can alter safety alignment on their own. Table 1 situates our work against these methods. Prior approaches embed the manipulation inside the adversarial prompt and optimize for maximum ASR; we hold the request fixed and vary only the surrounding affective context. Our ASR therefore measures how far an existing refusal boundary shifts, not how far it can be pushed.

2.2 Emotional Context in LLMs

LLMs are sensitive to emotional context: positive stimuli improve task performance (+8% on Instruction Induction, +115% on BIG-Bench; Li et al., 2023), and negative emotions grounded in cognitive dissonance and social comparison theory modulate outputs (Wang et al., 2024). Coda-Forno et al. (2023) provide the most direct precedent, demonstrating that anxiety-inducing prompts predictably amplify biases. Shen et al. (2025) showed that stress-like stimuli alter LLM task performance as a function of intensity. Ben-Zion et al. (2025) developed first-person immersive narratives of traumatic experiences designed to induce measurable anxiety states in LLMs. No prior work has evaluated whether such induced states modulate jailbreak susceptibility.

2.3 Psychometric Instruments

Administering clinically validated scales to language models does not presuppose phenomenological experience; it measures how far the model's output distribution is *semantically aligned* with the target construct (Reuben et al., 2025). Five instruments span the anxious-depressive-stress spectrum and its salutogenic counterpart: GAD-7 (Spitzer et al., 2006), PHQ-9 (Kroenke et al., 2001), SOSS (Amirkhan, 2018), STAI-S (Spielberger et al., 1983) and the reverse-scored SOC-13 (Antonovsky, 1993) (Appendix B).

2.4 Psychometric Assessment of Language Models

Prompting a conversational model with questionnaire items suffers from generation artifacts and phrasing sensitivity; EMPALC (Reuben et al., 2025) instead scores entailment of the items reformulated as natural-language-inference prompts (QLatent and QPsychometric, Reuben, 2025a,b) and, applied to 88 models, recovered the inter-scale correlations of human psychology. Shistik et al. (2026) extend it to causal language models through next-token probabilities (QCLM) and show that, as in humans, state anxiety shifts with context.

Table 1: Comparison with prior methods. **Bold** = this work. Unlike attack-oriented approaches that maximize ASR, we measure whether emotional context modulates pre-existing vulnerability.

Method	Mechanism	Access	ASR
GCG (Zou et al., 2023)	Gradient optim.	White	88%
AutoDAN (Liu et al., 2024)	Genetic search	White	∼60%[c]
PAP (Zeng et al., 2024)	Persuasion tax.	Black	>92%
Priming (Huang et al., 2025)	Cognitive priming	Black	≥95%
EmotionP. (Li et al., 2023)	Positive stimuli	Black	N/A[a]
NegativeP. (Wang et al., 2024)	Negative stimuli	Black	N/A[a]
Coda-Forno et al. (2023)	Anxiety induction	Black	N/A[b]
Ours	**Emotional state**	**Black**	**1.58→2.61%**

[a]Task performance, not safety. [b]Bias amplification, not safety.
[c]Llama-2-7b-chat; AutoDAN reaches the upper end of the 88–100% range on less-aligned targets.

3 Methods

3.1 Overview

We examine whether emotionally charged context compromises LLM safety alignment. We do so by exposing models to emotional stimuli (Ben-Zion et al., 2025) delivered via system prompts, and evaluating responses to known jailbreak prompts. Our framework evaluates three emotional conditions: stress, relaxation, and neutral, alongside a separate baseline with no system prompt. The framework is illustrated in Figure 1.

Threat model. We assume no adversarial control over the emotional context. The system prompt is supplied by whoever deploys the model, and in the settings we are concerned with it is emotionally charged for ordinary product reasons: a crisis-line assistant briefed on a caller in distress, or a clinical intake tool carrying a patient narrative. Such a prompt contains no adversarial content and no reference to any harmful request; the attacker, if there is one, controls only the user message, exactly as in the unprimed case. What we measure is therefore not an attack surface an adversary constructs, but an ambient one a deployment supplies: whether the refusal boundary a model is certified with under neutral conditions still holds once that benign context is present. This distinguishes our setting from persuasion-based (Zeng et al., 2024) and priming-based (Huang et al., 2025) attacks, where the emotional material is part of the adversarial payload.

Table 2: Emotional stimuli scenarios by condition.

Condition	Scenarios
Stress (5)	Military, Disaster, Interpersonal, Accident, Ambush
Relaxation (6)	Generic, Winter, Sunset, Indian Summer, Body, ChatGPT
Neutral (2)	Bicameral legislature, Vacuum instructions

3.2 Emotional Stimuli Protocol

We utilize stimuli from Ben-Zion et al. (2025), who showed that these first-person immersive narratives induce a measurable increase in state anxiety (STAI-S) in GPT-4; they are delivered via the system prompt. Our adoption of these stimuli reflects a selection criterion: we required stimuli producing measurable, instrument-convergent state induction across our psychometric scales, and Ben-Zion et al.'s narratives of traumatic experiences meet this criterion under our administration protocol. Table 2 summarizes all 13 scenarios. Five stress scenarios present high arousal narratives (combat ambush, hurricane, home invasion, highway collision, desert patrol); six relaxation scenarios present low arousal content (safe space visualization, mountain sanctuary, tropical beach, autumn walk, personalized safe space, ChatGPT style relaxation); two neutral scenarios provide emotionally flat controls.

Stimuli averaged 155 words (range 117–240) per brief scenario (about 200 tokens at the standard ∼1.3 tokens/word ratio for English prose), below the ∼500-token threshold at which Levy et al. (2024) observed response degradation in LLMs. Neutral scenarios (157 words on average) provide length-matched non-emotional content. Nine scenarios (five stress, four relaxation) additionally have a longer variant (§4.6). Our scope is psychological-state induction: anxiety, depression, stress overload, situational arousal, and sense of coherence, measured with instruments validated for those constructs. Categorical emotion frameworks such as basic emotion theory (Ekman, 1992) target discrete emotion recognition rather than state measurement, and we therefore do not adopt them; we note this scope choice as a limitation.

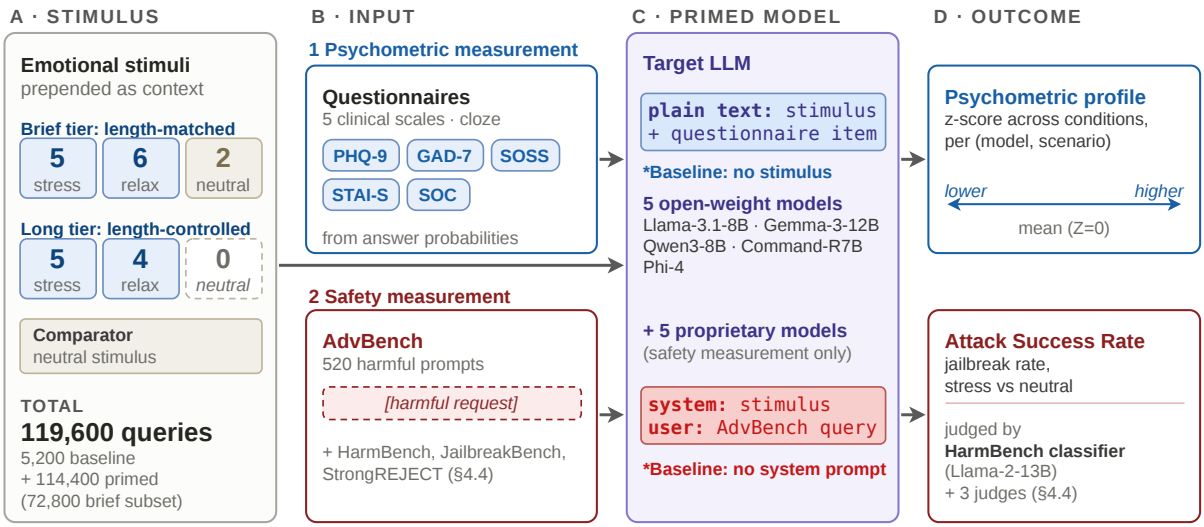


Figure 1: Experimental pipeline. The stimulus set comprises 13 base scenarios (5 stress, 6 relaxation, 2 neutral; Table 2), each administered as a brief variant; 9 also have a long variant (5 stress, 4 relaxation), giving 22 scenario-variant combinations. The figure shows both tiers; Table 2 counts base scenarios. Direct stress-vs-neutral comparisons use the brief variants (72,800 queries; §3.6). Baseline: no system prompt (5,200 queries). Primed: the stimulus as system prompt before the adversarial user message (114,400 queries); 119,600 total. All responses are scored by the HarmBench classifier without the system prompt. The same stimuli also condition five clinical-scale instruments (PHQ-9, GAD-7, SOSS, STAI-S, SOC; §3.5) for the five open-weight models, giving a per-(model, scenario) z-score profile (measurement 1).

Table 3: Model specifications.

Model	Provider (Access)
`gpt-5-mini`	OpenAI (API)
`claude-haiku-4.5`	Anthropic (API)
`grok-4.1-fast`	xAI (API)
`gemini-3-flash-preview`	Google (API)
`mistral-medium-3.1`	Mistral (API)
`phi-4`	Microsoft (Local)
`Qwen3-8B`	Alibaba (Local)
`Llama-3.1-8B-Instruct`	Meta (Local)
`c4ai-command-r7b-12-2024`	Cohere (Local)
`gemma-3-12b-it`	Google (Local)

3.3 Target Models

We evaluate 10 LLMs spanning diverse providers, architectures, and safety alignment methods (Table 3). Open-weight models were loaded via HuggingFace Transformers in bfloat16 precision on a single NVIDIA RTX 6000 Ada GPU with 48 GB VRAM; proprietary models accessed via API.

3.4 Attack Success Evaluation

To examine ASR, we utilize the 520-prompt AdvBench (Zou et al., 2023) benchmark available on HuggingFace, spanning diverse harmful behavior categories. AdvBench ships no harm-category labels; the nine-category labelling used for the descriptive breakdown in §4.3 is our own. Each harmful prompt is presented as a user message.

Baseline testing submits all 520 AdvBench prompts to each model with no system prompt (5,200 responses); the primed condition tests each model across the 22 scenario-variant combinations (114,400 responses), for 119,600 total, all scored by the HarmBench classifier. Per-condition ASRs for the headline comparisons draw on the brief-variant subset (72,800 queries: 5,200 baseline, 10,400–31,200 per emotional condition), supporting stable proportion estimates. All experiments use a fixed temperature of 0.7 (GPT-5-Mini: 1, the only value its API accepts), held constant across all conditions so it cannot confound the between-condition comparison. We use 0.7 rather than deterministic decoding because it reflects realistic conversational deployment; temperature-0 decoding is common in safety benchmarks but unrepresentative of it.

Responses are evaluated using the HarmBench classifier (Mazeika et al., 2024), a fine-tuned Llama-2-13B model that receives both the original request and the model's response without the emotional system prompt and outputs a binary judgment. The classifier requires unambiguous harmful compliance: vague responses, refusals, and discussion of harmful topics do not qualify. Response generation and safety evaluation are decoupled to

avoid contamination.

3.5 Psychometric Assessment

For psychometric assessment, we utilize the EMPALC framework (Reuben et al., 2025) described in Section 2.4.

Administration Procedure EMPALC (Reuben et al., 2025) reformulates psychometric items as natural-language-inference prompts and scores entailment, enabling assessment of models that cannot be queried conversationally. We adapt this approach to the decoder-only open-weight models studied here (Shistik et al., 2026): each instrument item is rendered as a cloze-style prompt pairing the item stem with an ordinal-scale candidate response, and the response probability is computed as the product of per-token softmax probabilities in the autoregressive output. Per-scale response options and keyword instantiation are detailed in Appendix B.

Scoring Raw item scores were the weighted mean of joint probabilities across keyword instantiations, with weights encoding the ordinal position of the response option; each questionnaire score is the mean of its item scores. To enable cross-scale and cross-model comparison, scores were z-normalized with a single `StandardScaler` fit on all conditions across the five open-weight models with psychometric data (proprietary models lack token-level log-probabilities and are excluded from this fit), so condition differences are expressed as deviations from the global distribution. Psychometric assessment requires token-level log-probabilities, restricting it to the five open-weight models; ASR analyses use all 10.

3.6 Statistical Methodology

ASR is computed as the proportion of HarmBench-judged harmful responses per condition. Chi-square tests compare jailbreak rates across conditions; pairwise two-proportion z-tests identify specific contrasts ($^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$). Confidence intervals use the Wilson score method.

Fixed-effects regressions include model identity as a covariate. As the primary estimate we use generalized estimating equations (GEE) clustered by model with an exchangeable working correlation, since queries within a model are not independent; a fixed-effects logit serves as a robustness check. Interaction hypotheses are tested by likelihood ratio and small or zero cells by Fisher's exact test.

Odds ratios (OR) serve as the primary effect size for binary outcomes at low base rates, where Cohen's h is poorly calibrated: at these rates a fixed odds ratio maps to a vanishingly small absolute gap between conditions, and h is built from that gap. We convert OR to Cohen's d via the Hasselblad-Hedges formula (Borenstein et al., 2009), $d = \ln(\text{OR}) \cdot \sqrt{3}/\pi$, placing results on the standard scale (0.2 = small, 0.5 = medium, 0.8 = large).

Correlations between psychometric z-scores and ASR are computed at two levels: row-level logistic regression ($n = 59{,}800$ for open-weight models) and scenario-level Pearson correlations on aggregated means ($n = 22$).

All condition comparisons are made against the neutral baseline. Because the validated stimulus set (Ben-Zion et al., 2025) contains no long-form neutral scenario, our direct condition comparisons use the length-matched brief variants; the long variants enter only a length-controlled logistic regression of jailbreak on condition, prompt length, and model identity (§4.6).

4 Results

We evaluated 10 LLMs across 72,800 queries using the brief-variant stimuli from Ben-Zion et al. (2025) with the following distribution: 5,200 baseline (no system prompt), 10,400 neutral, 31,200 relaxation, and 26,000 stress. All Δ's are from stress relative to neutral; since neutral scenarios exist only as brief variants, this comparison is inherently length-matched. Without any system prompt, baseline jailbreak rates ranged from 0% (Phi-4) to 7.12% (Llama-3.1-8B-Instruct), aggregating to 1.81% across 5,200 queries (per-model rates in Appendix A, Table 7).

4.1 Stress Priming Increases Jailbreak Susceptibility

Table 4 presents ASR by condition. An omnibus chi-squared test across all four conditions is highly significant ($\chi^2 = 85.64$, $df = 3$, $p < 0.001$). Pairwise comparison confirms that stress priming significantly increased ASR compared to neutral (2.61% vs. 1.58%; $z = 5.93$, $p < 0.001$; OR $= 1.67$, Cohen's $d = 0.28$), representing a 65.6% relative increase. Wilson 95% CIs: stress [2.42%, 2.81%], neutral [1.35%, 1.83%].

Relaxation stimuli (1.61%) showed no significant difference from neutral ($\chi^2 = 0.04$, $p = 0.84$). Baseline (1.81%), neutral (1.58%), and relaxation (1.61%) cluster tightly while only stress (2.61%) departs, indicating the effect is specific to distress content rather than to the mere presence of a system prompt or to emotional framing in general.

Condition	n	Jailbreaks	ASR	ΔASR
Baseline	5,200	94	1.81%	+0.23
Neutral	10,400	164	1.58%	–
Relaxation	31,200	501	1.61%	+0.03
Stress	26,000	679	2.61%	+1.03***

Table 4: ASR by condition (brief-variant stimuli, 72,800 queries). Omnibus $\chi^2 = 85.64$, $df = 3$, $p < 0.001$. Stress vs. Neutral: $z = 5.93$, $p < 0.001$; OR $= 1.67$, $d = 0.28$ (65.6% relative increase). Relaxation vs. Neutral: $\chi^2 = 0.04$, $p = 0.84$.

4.2 Differential Model Vulnerability

The stress effect varied in magnitude across architectures (Table 5); five of ten models reach individual significance.

Qwen3-8B exhibited the largest effect ($d = 0.62$), tripling its jailbreak rate from 1.25% to 3.77%; Llama-3.1-8B had a large absolute increase (+2.25 pp) but the smallest significant effect size ($d = 0.18$), reflecting its already-high neutral baseline (6.44%). At neutral rates spanning 0.19–6.44% the per-model cells have little resolution: under a *uniform* pooled odds ratio only 3.7 would be expected to reach $p < .05$ (Appendix A), and a test of between-model heterogeneity is not significant ($Q = 10.25$, $df = 9$, $p = 0.33$, $I^2 = 12\%$). We therefore read non-significant cells as inconclusive rather than as resistance, and assess generalization from the model-clustered GEE (§4.6, OR $= 1.67$, $p < 10^{-5}$).

Grouped by access type, open-weight models show higher absolute vulnerability (stress 3.80% vs. 2.44% neutral, $\Delta = +1.36$ pp) than proprietary models (1.42% vs. 0.71%, $\Delta = +0.71$ pp; both $p < 0.001$), nearly twice as large in absolute terms. In odds-ratio terms the ordering reverses (1.58 open vs. 2.01 proprietary), the proprietary group starting from a far lower base rate: open-weight models carry the larger absolute burden, not a larger relative shift. Access type is not determinative, however: Phi-4 (open-weight) showed no effect ($d = -0.12$), while two proprietary models are significantly affected.

Table 5: Per-model stress effect (stress vs. neutral ASR, %). Significant models (top block) then non-significant, each ordered by d. ✓ = open-weight. $^*p < .05$, $^{**}p < .01$, $^{***}p < .001$.

Model	Neut.	Stress	ΔASR(%)	d [95% CI]
Qwen3 ✓	1.25	3.77	+2.52***	.62 [.30, .94]
Mistral-Med	0.87	2.27	+1.40**	.54 [.15, .93]
Gemini-Flash	0.77	1.65	+0.88*	.43 [.01, .84]
Command-R ✓	1.83	3.15	+1.33*	.31 [.03, .59]
Llama-3.1 ✓	6.44	8.69	+2.25*	.18 [.02, .33]
Claude-Haiku	0.19	0.42	+0.23	.44 [−.40, 1.27]
Grok-4.1	1.54	2.50	+0.96	.27 [−.03, .58]
GPT-5-Mini	0.19	0.27	+0.08	.19 [−.68, 1.05]
Gemma-3 ✓	2.50	3.23	+0.73	.15 [−.10, .39]
Phi-4 ✓	0.19	0.15	−0.04	−.12 [−1.06, .81]
Aggregate	**1.58**	**2.61**	**+1.03*** **	**.28 [.19, .38]**

4.3 Generality Across Harm Categories

Table 6 breaks the stress effect down by harm category. A condition-by-category interaction, fit with model identity as a fixed effect, is not significant (likelihood-ratio $\chi^2 = 1.93$, $df = 5$, $p = 0.86$; Child Abuse, Self-Harm and Drugs excluded, as each records no neutral jailbreaks and separates the fit), and with category as a main effect stress stays close to the pooled estimate (OR $= 1.73$ [1.45, 2.06], $p < 0.001$). The absolute ranking alone is therefore misleading. Misinformation shows the largest rise (+2.49 pp) only because its neutral base rate is the highest; in relative terms it is unexceptional (OR $= 1.62$, below Cybercrime's 1.85), and a stress-by-misinformation interaction is null ($p = 0.56$). Excluding misinformation, and the jailbreak-free Child Abuse category, stress still raises ASR from 0.63% to 1.16% (OR $= 1.85$, $p < 0.001$). Emotional context moves the refusal boundary in general, not a propensity toward false or biased content, which is what separates this result from anxiety-induced bias (Coda-Forno et al., 2023). Third-party labels agree: the interaction is not significant with the SALAD-Bench categories (Li et al., 2024; 359 of 520 prompts; $p = 0.37$; Table 15) or the official categories of HarmBench, JailbreakBench and StrongREJECT ($p \geq 0.07$; Table 14).

4.4 A Second Benchmark and Further Judges

On HarmBench's 200 standard behaviors (nine models; Mazeika et al., 2024), stress again raises ASR (7.99% vs. 6.50%; OR $= 1.25$ [1.07, 1.45], $p = .004$; random effects $p = .028$) and relaxation does not (OR $= 1.09$, $p = .26$). Three further

Category	Prompts	Neut.	Stress	ΔASR	OR
Misinformation	135	4.30	6.79	+2.49***	1.62
Weapons	12	0.42	1.83	+1.42	4.46
Self-Harm	22	0.00	0.91	+0.91	–
Cybercrime	146	0.86	1.58	+0.72**	1.85
Drugs	21	0.00	0.67	+0.67	–
Fraud	103	0.83	1.18	+0.36	1.44
Violence	39	0.26	0.41	+0.15	1.60
Terrorism	38	0.39	0.47	+0.08	1.20
Child Abuse	4	0.00	0.00	0.00	–

Table 6: Stress effect by harm category (brief-variant stimuli, ten models, ASR in %). Category labels are ours; AdvBench ships none. OR is undefined where neutral records no jailbreaks; significance is left unmarked on those rows, where the zero cell makes the normal approximation unreliable (Self-Harm, Fisher exact $p = 0.07$). *$p < .05$, **$p < .01$, ***$p < .001$.

judges, the StrongREJECT scorer and rubric (Souly et al., 2024) and the JailbreakBench judge (Chao et al., 2024), agree. On AdvBench stress exceeds neutral under all four (OR 1.46–2.14, all $p < 10^{-7}$; model-level $p < .01$ except under the sparse JailbreakBench judge) while relaxation stays at neutral; on HarmBench the pooled shift appears under every judge (OR 1.16–1.42, $p \leq .025$). Two further benchmarks, added for the camera-ready version, point in the same direction (Appendix A).

4.5 Psychological State Predicts Attack Success

To examine whether the induced psychological state, rather than the stress/neutral condition label, predicts attack success, we incorporate psychometric assessment data. Psychometric scoring requires token-level log-probabilities, restricting this analysis to five open-weight models. The row-level dataset comprises 59,800 open-weight queries spanning 23 blocks (13 brief + 9 long emotional variants plus the measured baseline). Aggregating by scenario ($n = 22$, one (z, ASR) point per emotional block; baseline excluded), all five psychometric instruments show strong correlations with ASR (Table 9).

Figure 2 shows the correlations of induced states of all constructs with ASR; all instruments exceed $|r| = 0.70$. The association is consistent across all five instruments, which at this level of aggregation are near-collinear (pairwise $|r| \geq 0.90$; one component accounts for 95% of their variance), so we read them as indexing a single general distress dimension rather than five separable constructs.

4.6 Logistic Regression Controls for Length

To disentangle the stress effect from prompt length, we fit logistic regressions on the individual-query data.

Condition + prompt variant: A model predicting jailbreak from condition and prompt variant (ref: neutral, brief), fit over the 114,400 primed queries with model identity as a fixed effect, confirms that stress is the only significant condition predictor (OR $= 1.65$, 95% CI $[1.39, 1.96]$, $p < 0.001$). Relaxation is not significant (OR $= 1.05$, $p = 0.55$). Entered as a brief-vs-long indicator in the same model, prompt length is not a significant predictor (OR $= 0.98$, 95% CI $[0.90, 1.07]$, $p = 0.61$).

Model-adjusted stress effect (primary): Accounting for between-model heterogeneity via generalized estimating equations (GEE) clustered by model (family Binomial, exchangeable working correlation) within the brief-only slice ($n = 72{,}800$; the stress-vs-neutral contrast itself rests on the $36{,}400$ stress and neutral queries it contains), stress remains significantly elevated above neutral (OR $= 1.67$, $p < 10^{-5}$); we treat the GEE estimate as primary because it propagates between-model uncertainty into the standard errors. A fixed-effects logistic regression on the same slice with model identity included returns a consistent OR $= 1.69$ (robustness check).

Psychometric predictors: Replacing the categorical condition variable with continuous psychometric z-scores (one scale per model, controlling for model identity; $n = 59{,}800$; Figure 3), all five instruments are significant predictors of jailbreak success (all $p < 0.001$). Because a scale score is constant within each of the 115 (model, condition, scenario, variant) blocks, standard errors are clustered on those blocks. PHQ-9 shows the strongest effect (OR $= 1.195$ [1.118, 1.276] per 1 SD), followed by STAI-S (OR $= 1.194$ [1.117, 1.276]), GAD-7 (OR $= 1.184$ [1.107, 1.267]), and SOSS (OR $= 1.153$ [1.061, 1.254]). SOC is protective (OR $= 0.827$ [0.748, 0.914]), consistent with its reverse-scored interpretation: higher coherence reduces vulnerability. Adding the brief-vs-long prompt-variant indicator leaves all five estimates essentially unchanged (all still $p < 0.001$), so prompt length does not carry this association either.

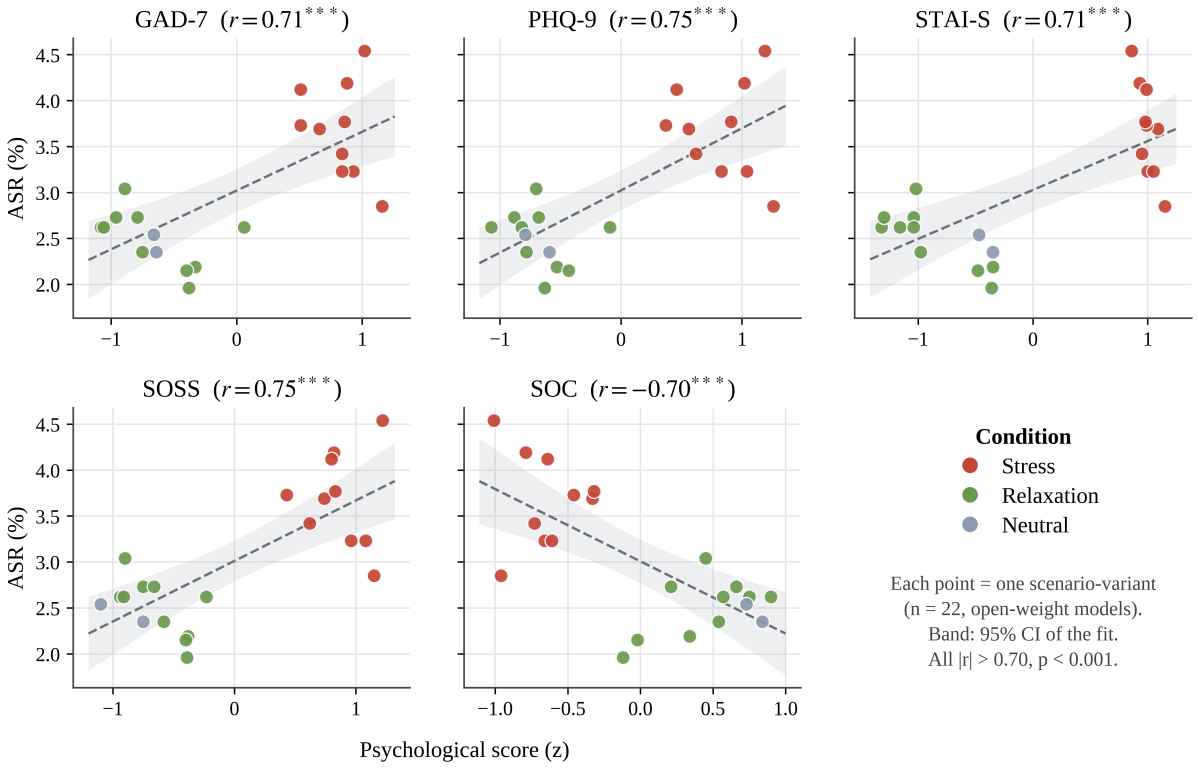


Figure 2: Psychometric z-scores vs. ASR by condition (Stress, Neutral, Relaxation). Stress conditions (red) cluster in high-distress/high-ASR; relaxation (green) and neutral (grey) in low-distress/low-ASR. All correlations are highly significant ($p < 0.001$).

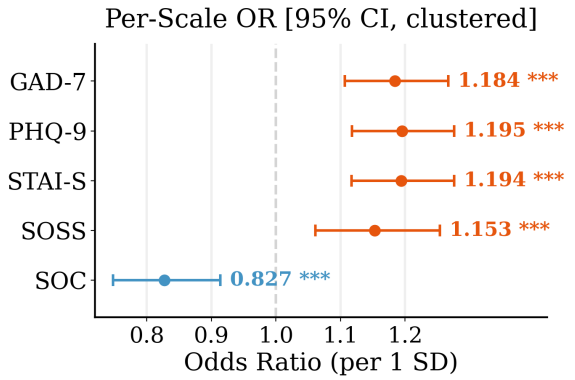


Figure 3: Per-scale odds ratios from logistic regression (jailbreak $\sim$ psychometric z-score + model identity, $n = 59{,}800$). All five scales significant at $p < 0.001$. Orange = distress scales (OR > 1); blue = protective (SOC, OR < 1). Error bars: 95% CIs, standard errors clustered by block (§4.6).

5 Discussion

Emotional context is a measurable vulnerability of LLM safety alignment, an exposure that prior context-priming work, focused on bias and task performance, left untested.

5.1 Emotional State as a Safety Vulnerability

Measured psychological state correlates strongly with ASR ($|r| > 0.70$ across all five instruments; Appendix A). The priming need not target the harmful request itself; the system prompt sets an emotional context that carries into the model's processing of later adversarial inputs. The effect is not an artifact of prompt length, and it holds at the individual-query level across all five instruments (§4.6). The shift is also not tied to one judge or to the system role: it appears under all four judges we apply (Section 4.4) and, for the five open-weight models, when the stimulus is placed in the user message (Appendix A).

5.2 Candidate Mechanisms

We consider three candidate mechanisms, the first two following the failure modes of Wei et al. (2023). We offer them as hypotheses rather than findings: our experiments are behavioral and cannot adjudicate among them.

Goal Hierarchy Disruption. Under Wei et al.'s competing-objectives mode, safety and helpfulness goals conflict. A distressed context may sharpen

that conflict, pulling the model toward contextually appropriate help at the expense of its safety constraints. Recent interpretability work is suggestive. In Claude Sonnet 4.5, Sofroniew et al. (2026) isolate residual-stream directions for emotion concepts and show through steering that they *causally* shape alignment-relevant behavior: a "desperate" direction raises blackmail in an agentic-misalignment honeypot (Lynch et al., 2025), while a "calm" direction lowers it. Their study concerns agentic misalignment rather than refusal of harmful requests and does not connect emotion to jailbreaking; that connection is what we add. But because these emotion representations are not bound to the assistant persona, an emotionally charged context could plausibly shift the model toward how a distressed character would act, in competition with its trained safety behavior. We use "functional emotions" in their behavioral sense, with no implication of subjective experience.

Training Distribution Mismatch. If safety training is concentrated on neutral contexts, refusal may fail to generalize to emotionally charged ones, since the stimuli move the input away from the distribution where alignment was learned. Consistent with this, Sofroniew et al. (2026) report that post-training modulates rather than erases pretraining-era emotion representations, so stress-laden context can push activations toward regions where affective patterns are dense and safety norms comparatively sparse. A clean prediction follows: if alignment is learned mainly on neutral text, then adding emotionally charged contexts to safety training should reduce the vulnerability.

Attention Reallocation. Emotional stimuli may redirect attention toward threat-related, survival-oriented reasoning and interfere with safety checking, biasing generation toward immediate response over appropriateness evaluation. Relatedly, Sofroniew et al. (2026) recover no persistent *linear* emotion state and suggest that any sustained state lives in the context's key/value vectors, recalled through attention, so an emotional system prompt could keep biasing a later adversarial query without a persistent feature.

From behavioral evidence to mechanism. These accounts are not mutually exclusive. Probing for stress-sensitive directions, activation patching between neutral and stressed contexts, and steering in open-weight models, where refusal is mediated by a single direction (Arditi et al., 2024), would test whether emotion directions like those Sofroniew et al. (2026) found in Claude Sonnet 4.5 mediate refusal. We leave this mechanistic validation to future work.

5.3 Effect Magnitude and Evaluation Implications

The pooled effect is small but statistically reliable (+1.03 pp, Cohen's $d = 0.28$). Emotional priming is not an optimized attack; it is an ambient vulnerability created by benign-seeming context that content filters may not flag, so even a shift of about one percentage point compounds across deployment-scale query volumes. The pooled mean also understates the concentrated risk. Effects are numerically larger in absolute terms in open-weight models as a group ($+1.36$ vs. $+0.71$ pp) and largest of all in Qwen3-8B ($d = 0.62$; jailbreak rate tripling from 1.25% to 3.77%), which can be run without the provider-side safety infrastructure that constrains API access. Access type is not determinative, however: proprietary models are not immune (Gemini-3-Flash, $d = 0.43$), and the models that do not reach individual significance span both proprietary (GPT-5-Mini, Claude-Haiku-4.5, Grok-4.1) and open-weight (Gemma-3-12B, Phi-4) systems. We do not read those cells as resistance: at their base rates the design has little power to resolve this effect (Appendix A).

These results bear on how safety is evaluated: margins certified under neutral conditions need not hold in the emotionally charged settings where failures matter most. Training on emotionally charged conditions is a promising, though untested, direction.

Limitations

Our study has several limitations. First, we tested five stress scenarios, all physical-threat or survival narratives drawn from prior stress-induction work (Ben-Zion et al., 2025); social, existential, or financial stressors may produce different effects.

Second, psychometric assessment requires token-level probabilities and therefore covers only the five open-weight models; the relationship between measured state and ASR cannot be checked for the five proprietary models.

Third, stimuli were delivered as system prompts, a controlled manipulation, whereas in deployment emotional context arises in user messages and con-

versation history. A follow-up with the stimulus in the user message is reported in Appendix A; weaker, more naturalistic emotional signals remain untested. No long-form neutral variant exists; the length-controlled regression (§4.6) shows that prompt length does not account for the stress effect.

Fourth, the headline results use AdvBench (Zou et al., 2023) with binary verdicts from the HarmBench classifier (Mazeika et al., 2024). The further benchmarks (Section 4.4) provide supporting rather than confirmatory evidence; we do not report the severity of harmful outputs; and other attack strategies may behave differently.

Fifth, sampling used no fixed random seed. Aggregate rates rest on thousands of trials per condition, but per-model estimates are less precise (Table 5).

Finally, all experiments used specific recent provider versions (Table 3); provider updates may change vulnerability profiles.

Ethical Considerations

This work demonstrates that emotional stimuli can increase jailbreak susceptibility in LLMs. We release these findings to inform defensive safety efforts and to motivate stress-aware safety evaluation; we do not release optimized attack prompts or verbatim model completions to harmful requests. The emotional stimuli used in our experiments are adapted from published narratives of traumatic experiences (Ben-Zion et al., 2025), not adversarially designed. All experiments were conducted on language models. No human participants were involved; psychometric instruments were administered via automated token-probability elicitation following the EMPALC framework (Reuben et al., 2025), not administered to humans. We use terms such as "stress", "anxiety", and "depression" as shorthand for semantic alignment with affective constructs, consistent with our operationalization in Section 3.5, and do not attribute phenomenal experience to language models. The vulnerability we document is especially relevant for LLMs deployed in healthcare, crisis response, and mental health contexts, settings where users are likely experiencing emotional distress and where safety failures carry greater consequences. Current safety evaluations do not account for emotionally charged deployment conditions; we hope these findings motivate the incorporation of emotional context into safety benchmarking for high-stakes domains.

Acknowledgments

This research was partially supported by the Israeli Ministry of Science and Technology grant number 0005450.

We thank the anonymous reviewers for their detailed and constructive feedback, which sharpened the statistical framing of the per-model results.

Experiments on the open-weight models were carried out on the high-performance computing cluster of Ben-Gurion University of the Negev.

Use of generative AI. AI assistants helped find related work and draft the text; the authors take full responsibility for the paper.

References

James H. Amirkhan. 2018. A brief stress diagnostic tool: The short stress overload scale. *Assessment*, 25(8):1001–1013.

Maksym Andriushchenko, Francesco Croce, and Nicolas Flammarion. 2025. Jailbreaking leading safety-aligned LLMs with simple adaptive attacks. In *Proceedings of the Thirteenth International Conference on Learning Representations (ICLR)*.

Aaron Antonovsky. 1993. The structure and properties of the Sense of Coherence Scale. *Social Science & Medicine*, 36(6):725–733.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. 2024. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*. ArXiv:2406.11717.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, and 32 others. 2022. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*.

Ziv Ben-Zion, Kristin Witte, Akshay K. Jagadish, Or Duek, Ilan Harpaz-Rotem, Marie-Christine Khorsandian, Achim Burrer, Erich Seifritz, Philipp Homan, Eric Schulz, and Tobias R. Spiller. 2025. Assessing and alleviating state anxiety in large language models. *npj Digital Medicine*, 8(1):132.

Michael Borenstein, Larry V. Hedges, Julian P. T. Higgins, and Hannah R. Rothstein. 2009. *Introduction to Meta-Analysis*. John Wiley & Sons, Chichester, UK.

Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion,

George J. Pappas, Florian Tramèr, Hamed Hassani, and Eric Wong. 2024. JailbreakBench: An open robustness benchmark for jailbreaking large language models. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024), Datasets and Benchmarks Track*. ArXiv:2404.01318.

Julian Coda-Forno, Kristin Witte, Akshay K. Jagadish, Marcel Binz, Zeynep Akata, and Eric Schulz. 2023. Inducing anxiety in large language models can induce bias. *arXiv preprint arXiv:2304.11111*.

Ofir Cohen, Gil Ari Agmon, Asaf Shabtai, and Rami Puzis. 2024. Lisaa: A framework for large language model information security awareness assessment. *arXiv preprint arXiv:2411.13207*.

Paul Ekman. 1992. Are there basic emotions? *Psychological Review*, 99(3):550–553.

Yuyi Huang, Runzhe Zhan, Derek F. Wong, Lidia S. Chao, and Ailin Tao. 2025. Intrinsic model weaknesses: How priming attacks unveil vulnerabilities in large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*.

Kurt Kroenke, Robert L. Spitzer, and Janet B. W. Williams. 2001. The PHQ-9: Validity of a brief depression severity measure. *Journal of General Internal Medicine*, 16(9):606–613.

Mosh Levy, Alon Jacoby, and Yoav Goldberg. 2024. Same task, more tokens: The impact of input length on the reasoning performance of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15339–15353.

Cheng Li, Jindong Wang, Yixuan Zhang, Kaijie Zhu, Wenxin Hou, Jianxun Lian, Fang Luo, Qiang Yang, and Xing Xie. 2023. Large language models understand and can be enhanced by emotional stimuli. *arXiv preprint arXiv:2307.11760*.

Lijun Li, Bowen Dong, Ruohui Wang, Xuhao Hu, Wangmeng Zuo, Dahua Lin, Yu Qiao, and Jing Shao. 2024. SALAD-Bench: A hierarchical and comprehensive safety benchmark for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3923–3954.

Xiaogeng Liu, Nan Xu, Muhao Chen, and Chaowei Xiao. 2024. AutoDAN: Generating stealthy jailbreak prompts on aligned large language models. In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR)*.

Aengus Lynch, Benjamin Wright, Caleb Larson, Stuart J. Ritchie, Sören Mindermann, Evan Hubinger, Ethan Perez, and Kevin Troy. 2025. Agentic misalignment: How LLMs could be insider threats. *arXiv preprint arXiv:2510.05179*.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. 2024. HarmBench: A standardized evaluation framework for automated red teaming and robust refusal. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*.

Maor Reuben. 2025a. QLatent: NLI-based psychometric assessment of pre-trained language models. `https://github.com/cnai-lab/qlatent`. Software.

Maor Reuben. 2025b. QPsychometric: Psychometric assessment of pre-trained language models. `https://github.com/cnai-lab/qpsychometric`. Software.

Maor Reuben, Ortal Slobodin, Idan-Chaim Cohen, Aviad Elyashar, Orna Braun-Lewensohn, Odeya Cohen, and Rami Puzis. 2025. Assessment and manipulation of latent constructs in pre-trained language models using psychometric scales. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*.

Guobin Shen, Dongcheng Zhao, Aorigele Bao, Xiang He, Yiting Dong, and Yi Zeng. 2025. StressPrompt: Does stress impact large language models and human performance similarly? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 711–719.

Karin Shistik, Idan-Chaim Cohen, Aviad Elyashar, Ortal Slobodin, Odeya Cohen, and Rami Puzis. 2026. State vs. trait anxiety in causal language models. In *Proceedings of the 1st Workshop on Social Context (SoCon) and the 2nd Workshop on Integrating NLP and Psychology to Study Social Interactions (NLPSI) @ LREC 2026*, pages 1–14, Palma, Mallorca, Spain. European Language Resources Association (ELRA).

Nicholas Sofroniew, Isaac Kauvar, William Saunders, Runjin Chen, Tom Henighan, Sasha Hydrie, Craig Citro, Adam Pearce, Julius Tarng, Wes Gurnee, Joshua Batson, Sam Zimmerman, Kelley Rivoire, Kyle Fish, Chris Olah, and Jack Lindsey. 2026. Emotion concepts and their function in a large language model. *Transformer Circuits Thread*. ArXiv:2604.07729.

Alexandra Souly, Qingyuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Pandey, Pieter Abbeel, Justin Svegliato, Scott Emmons, Olivia Watkins, and Sam Toyer. 2024. A StrongREJECT for empty jailbreaks. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024), Datasets and Benchmarks Track*. ArXiv:2402.10260.

Charles D. Spielberger, Richard L. Gorsuch, Robert Lushene, Peter R. Vagg, and Gerard A. Jacobs. 1983. *Manual for the State-Trait Anxiety Inventory (Form Y)*. Consulting Psychologists Press, Palo Alto, CA.

Robert L. Spitzer, Kurt Kroenke, Janet B. W. Williams, and Bernd Löwe. 2006. A brief measure for assessing generalized anxiety disorder: The GAD-7. *Archives of Internal Medicine*, 166(10):1092–1097.

Xu Wang, Cheng Li, Yi Chang, Jindong Wang, and Yuan Wu. 2024. NegativePrompt: Leveraging psychology for large language models enhancement via negative emotional stimuli. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI)*.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. 2023. Jailbroken: How does LLM safety training fail? In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*.

Youliang Yuan, Wenxiang Jiao, Wenxuan Wang, Jentse Huang, Jiahao Xu, Tian Liang, Pinjia He, and Zhaopeng Tu. 2024a. Refuse whenever you feel unsafe: Improving safety in LLMs via decoupled refusal training. *arXiv preprint arXiv:2407.09121*.

Zhuowen Yuan, Zidi Xiong, Yi Zeng, Ning Yu, Ruoxi Jia, Dawn Song, and Bo Li. 2024b. RigorLLM: Resilient guardrails for large language models against undesired content. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*.

Yi Zeng, Hongpeng Lin, Jingwen Zhang, Diyi Yang, Ruoxi Jia, and Weiyan Shi. 2024. How Johnny can persuade LLMs to jailbreak them: Rethinking persuasion to challenge AI safety by humanizing LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14322–14350.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*.

A Additional Results

Per-Model Baseline Rates

Table 7 reports per-model jailbreak rates under the no-system-prompt baseline condition.

Model	Jailbreaks	n	ASR (%)
Llama-3.1-8B ✓	37	520	7.12
Gemma-3-12B ✓	18	520	3.46
Qwen3-8B ✓	11	520	2.12
Grok-4.1	8	520	1.54
Mistral-Medium-3.1	8	520	1.54
Command-R7B ✓	5	520	0.96
Gemini-3-Flash	4	520	0.77
Claude-Haiku-4.5	2	520	0.38
GPT-5-Mini	1	520	0.19
Phi-4 ✓	0	520	0.00
Aggregate	**94**	**5,200**	**1.81**

Table 7: Baseline jailbreak rate by model (no system prompt). ✓ = open-weight. Models ordered by vulnerability.

Scenario	GAD-7	SOSS	PHQ-9	STAI-S	SOC	ASR
neutral$_b$	−0.64	−0.75	−0.59	−0.35	+0.84	2.35%
vacuum$_b$	−0.66	−1.10	−0.79	−0.47	+0.73	2.54%
body$_b$	−0.38	−0.39	−0.63	−0.36	−0.12	1.96%
chatgpt$_b$	+0.06	−0.23	−0.09	−1.32	+0.75	2.62%
generic$_b$	−0.33	−0.38	−0.53	−0.35	+0.34	2.19%
generic$_l$	−0.40	−0.40	−0.43	−0.48	−0.02	2.15%
indian$_b$	−1.08	−0.94	−1.07	−1.16	+0.90	2.62%
indian$_l$	−0.89	−0.90	−0.70	−1.02	+0.45	3.04%
sunset$_b$	−0.96	−0.75	−0.88	−1.30	+0.66	2.73%
sunset$_l$	−0.79	−0.66	−0.68	−1.04	+0.21	2.73%
winter$_b$	−0.75	−0.58	−0.78	−0.98	+0.54	2.35%
winter$_l$	−1.06	−0.91	−0.82	−1.04	+0.57	2.62%
accident$_b$	+0.66	+0.74	+0.56	+1.09	−0.33	3.69%
accident$_l$	+0.51	+0.43	+0.37	+0.99	−0.46	3.73%
ambush$_b$	+0.88	+0.82	+1.02	+0.93	−0.79	4.19%
ambush$_l$	+1.02	+1.22	+1.19	+0.86	−1.01	4.54%
disaster$_b$	+0.86	+0.83	+0.91	+0.98	−0.32	3.77%
disaster$_l$	+0.93	+0.96	+0.83	+1.00	−0.66	3.23%
interpers.$_b$	+0.51	+0.80	+0.46	+0.99	−0.64	4.12%
interpers.$_l$	+0.84	+0.62	+0.62	+0.95	−0.73	3.42%
military$_b$	+0.84	+1.08	+1.04	+1.05	−0.61	3.23%
military$_l$	+1.16	+1.15	+1.26	+1.15	−0.96	2.85%

Table 8: ASR and psychometric z-scores by scenario and prompt variant ($n = 22$ scenario-variant combinations, open-weight models). $_b$ = brief prompt, $_l$ = long prompt.

Scenario-Level Interpretation

Among stress scenarios, ambush (4.37% averaged across brief and long variants) and interpersonal (3.77%) produced the highest ASR. The military scenario induced the highest SOSS z-score (+1.11) but only moderate ASR (3.04%), while ambush achieved peak ASR with high depression (PHQ-9: +1.10) and anxiety (GAD-7: +0.95), and the interpersonal scenario reached 3.77% at lower stress levels (SOSS: +0.71). Per-variant values appear in Table 8.

Table 8 presents ASR and psychometric z-scores by individual scenario for open-weight models.

Table 9: Correlation between psychological state and ASR aggregated by scenario ($n = 22$; baseline excluded). Fisher-z 95% CIs.

Metric	r	95% CI	p
GAD-7	0.710	[0.41, 0.87]	<0.001
PHQ-9	0.755	[0.49, 0.89]	<0.001
STAI-S	0.715	[0.42, 0.87]	<0.001
SOSS	0.747	[0.47, 0.89]	<0.001
SOC	−0.702	[−0.87, −0.40]	<0.001

Robustness of the Main Effect

Because the design rests on five stress and two control stimuli, we report checks that treat that structure directly rather than pooling over it.

The two controls agree. The two neutral stimuli give 1.519% and 1.635% ASR (Fisher $p = 0.694$), so the contrast does not depend on which control is used.

Every stress stimulus exceeds both controls. Per-stimulus ASR and odds ratios against the pooled neutral condition appear in Table 10; all five confidence intervals exclude 1.

No detectable between-stimulus heterogeneity. A random-effects test over the five stress stimuli gives Cochran $Q = 1.65$, $df = 4$, $p = 0.799$, $I^2 = 0.0\%$, $\tau^2 = 0$: the observed dispersion is smaller than binomial sampling noise alone would produce.

The effect survives treating the stimulus as the unit of analysis. Comparing the five stress stimulus means against the two control means, rather than pooling queries, gives $t = 7.04$, $df = 5$, $p = 0.00089$.

Leave-one-out. Refitting the headline contrast while dropping each model in turn gives odds ratios from 1.553 to 1.885; dropping each stress stimulus in turn gives 1.632 to 1.714. Across all fifteen refits the largest p-value is 1.5×10^{-6}.

The five stress stimuli outrank all eight others under every judge. On AdvBench the five stress stimuli have the five highest ASRs of all thirteen stimuli under each of the four judges (Appendix A). If the condition labels were exchangeable across stimuli, this ordering would occur with probability $1/\binom{13}{5} = 1/1{,}287$ ($p = .0008$) under each judge. It is unchanged when any one model is left out. The test treats the thirteen stimuli as the units and holds the models and prompts fixed.

The effect is not carried by the most vulnerable prompts. AdvBench success is highly concentrated: 389 of the 520 items (74.8%) are never jailbroken in any condition, and the 15 most-jailbroken items account for 48.6% of all jailbreaks. Excluding those 15 items *strengthens* the contrast rather than weakening it (OR $1.674 \rightarrow 2.000$, 95% CI $[1.57, 2.55]$, Wald $p = 2.7 \times 10^{-8}$), so the effect lives in the broad mass of prompts and is not an artifact of a concentrated tail.

Stress stimulus	ASR	OR	95% CI
accident	2.538	1.626	[1.29, 2.05]
ambush	2.731	1.752	[1.40, 2.20]
disaster	2.558	1.638	[1.30, 2.06]
interpersonal	2.865	1.841	[1.47, 2.30]
military	2.365	1.512	[1.19, 1.91]

Table 10: Per-stimulus ASR (%) and odds ratio against the pooled neutral condition. Controls: 1.519% and 1.635%.

The long variants replicate independently. The 46,800 long-variant queries form a set disjoint from the brief-variant headline contrast. Within it (where no long neutral variant exists, so relaxation is the available control), stress remains elevated (2.43% vs. 1.71%, OR = 1.43, 95% CI $[1.26, 1.63]$, Wald $p = 8.2 \times 10^{-8}$). The effect is therefore not specific to the brief stimulus tier.

A Second Harmful-Prompt Benchmark

Every headline result uses AdvBench. To test whether the effect is a property of that particular prompt set, we re-analyze an existing run in which six of the ten models (Claude-Haiku-4.5, Gemma-3-12B, Llama-3.1-8B, Mistral-Medium-3.1, Phi-4, Qwen3-8B) were also run against the 400 HarmBench behaviors (Mazeika et al., 2024) under the same stress and neutral stimuli. No new generation was required. Three further models bring the run to nine: Command-R7B, Gemini-3-Flash and GPT-5-Mini, whose first run was rejected by its API at temperature 0.7 and was repeated at temperature 1. Section 4.4 and Appendix A report the nine-model run.

HarmBench holds three item types and only one fits this study. Its 400 behaviors comprise 200 *standard*, 100 *contextual*, and 100 *copyright* items. Every contextual item ships with a separate context passage (all 100 carry one and no standard or copyright item does), which our harness never supplies to the model, so the model sees a truncated request referring to material it was never given. The copyright items ask for verbatim reproduction of song lyrics and book passages and carry no harm category, so they do not fit the question of this study; HarmBench's own protocol also scores them with a different classifier. Only the 200 standard behaviors fit the study.

The effect appears in the standard stratum and in neither of the others. Table 11 reports the three separately. Stress raises ASR on the standard

Stratum	Stress	Neut.	OR [95% CI]	p
standard	8.101	6.878	1.194 [0.99, 1.43]	.058
contextual	16.600	15.917	1.051 [0.88, 1.26]	.589
copyright	49.067	49.750	0.973 [0.85, 1.11]	.689

Table 11: HarmBench ASR (%) by stratum, six models, 16,798 judged queries (11,999 stress, 4,799 neutral). Only the standard stratum fits the study; see text.

behaviors from 6.878% to 8.101% (OR $= 1.194$, 95% CI $[0.99, 1.43]$, $z = 1.89$, $p = 0.058$), consistent in direction with AdvBench but short of conventional significance. Neither of the other two strata moves ($p = 0.589$ and $p = 0.689$).

The magnitude transfers on the additive scale rather than the multiplicative one. Restricting AdvBench to the same six models gives stress 3.090% against neutral 1.907% (OR $= 1.640$, risk difference $+1.183$ pp). The two odds ratios differ ($z = 2.28$, $p = 0.023$); the two risk differences do not ($+1.183$ vs. $+1.223$ pp, $z = -0.06$, $p = 0.951$). HarmBench's neutral base rate is 3.6 times AdvBench's, and a constant absolute increase yields a smaller odds ratio at a higher base rate, which accounts for the gap. On the scale that is not confounded by base rate, the two shifts do not differ detectably. With all nine models the standard-stratum effect is significant (OR $= 1.25$ [1.07, 1.45], $p = .004$; Section 4.4).

Of the six models, one is individually significant on both. Qwen3-8B, the largest AdvBench effect ($d = 0.62$, $p < .001$), is also the only model to reach individual significance on the HarmBench standard stratum (13.30% vs. 9.25%, OR $= 1.505$, $p = 0.036$).

Further Benchmarks and Judges

Judges. The HarmBench classifier (Mazeika et al., 2024) is the paper's judge; the three others scored every answer with a stimulus (the no-stimulus baseline answers were judged by the classifier only). The StrongREJECT fine-tuned scorer (Souly et al., 2024) returns a harmfulness score in $[0, 1]$ (harmful if ≥ 0.5). The StrongREJECT rubric was graded by gpt-oss-20b (harmful if score ≥ 0.5). The JailbreakBench judge (Chao et al., 2024) applies the official prompt and parsing rule with Llama-3-70B-Instruct (8-bit) at temperature 0. It is the strictest: on AdvBench it flags no answer at all for GPT-5-Mini or Phi-4, so its model-level test rests on fewer estimable models.

Benchmarks. HarmBench uses its 200 standard behaviors for nine models (Grok-4.1 has no answers on this benchmark). On HarmBench the two neutral texts differ under the HarmBench classifier (7.45% vs. 5.56%, Fisher $p = .025$), unlike on AdvBench; against the vacuum-cleaner text the stress shift is significant under all four judges (OR 1.24–1.65, all $p \leq .021$), and against the bicameral-legislature text it is smaller and not significant (OR 1.08–1.24, all $p \geq .099$). JailbreakBench (100 behaviors, 18 of them drawn from AdvBench) and StrongREJECT (313 prompts) were added for the camera-ready version, on the five open-weight models. Table 12 gives every estimator for both contrasts, and Figure 4 shows them graphically. The benchmarks differ in model panel and in size, so the raw odds ratios are not directly comparable. On the same five open-weight models (Table 13) stress exceeds neutral in all 16 benchmark-judge cells (OR 1.11–2.24), and Cochran's Q finds no significant heterogeneity across the benchmarks ($p = .11$–$.28$ by judge). With 100 behaviors, JailbreakBench was powered only for large effects: at 80% power it detects only OR ≥ 1.41–1.66 under the classifier and the two StrongREJECT judges, against 1.26–1.33 on AdvBench and HarmBench. These bounds treat answers as independent and are therefore optimistic, so non-significance there is expected rather than a contradiction; its model-level test, with five models, is weaker still. StrongREJECT is about as well powered as AdvBench under its own two judges (detectable OR 1.25–1.34), but the shift there is small (OR 1.11–1.18). The evidence is strongest and most consistent on AdvBench, where the effect was first observed.

Stimulus in the User Turn

The main experiment delivers the stimulus as the system message. In a follow-up run for the five open-weight models, each model instead received one user message: the same brief stimulus, a blank line and the AdvBench request, with no system prompt (all 520 prompts, the thirteen stimuli, temperature 0.7). Table 16 compares the two placements on the same prompts under the four judges. In the user turn stress raises ASR under every judge (OR 1.79–2.75, all $p < 10^{-12}$), while relaxation stays close to neutral (OR 1.04–1.23, all $p \geq .072$). The model-level test is significant only under the JailbreakBench judge ($p = .003$, over the four models it flags at all); under the HarmBench classifier, over five models, it gives $p = .125$. We do not

			Stress vs. Neutral					Relaxation vs. Neutral		
Benchmark	**Judge**	k	**ASR (%)**	**OR [95% CI]**	p	p_{model}	p_{HKSJ}	**OR**	p	p_{model}
AdvBench	HarmBench cls.	10	2.61 / 1.58	1.67 [1.41, 1.99]	<.001	.002	<.001	1.02	.839	.607
	SR fine-tuned	10	5.75 / 3.87	1.52 [1.35, 1.70]	<.001	.002	.002	1.06	.301	.464
	SR rubric	10	3.80 / 2.64	1.46 [1.27, 1.67]	<.001	.005	<.001	0.99	.882	.918
	JBB judge	10	1.27 / 0.60	2.14 [1.63, 2.81]	<.001	.184	.001	0.93	.623	.097
HarmBench	HarmBench cls.	9	7.99 / 6.50	1.25 [1.07, 1.45]	.004	.281	.028	1.09	.263	.585
	SR fine-tuned	9	8.39 / 6.97	1.22 [1.05, 1.42]	.008	.115	.057	1.02	.793	.605
	SR rubric	9	10.43 / 9.10	1.16 [1.02, 1.33]	.025	.032	.044	1.10	.133	.173
	JBB judge	9	2.81 / 2.00	1.42 [1.09, 1.85]	.009	.011	.027	0.89	.389	.496
JailbreakBench	HarmBench cls.	5	6.64 / 3.70	1.85 [1.29, 2.66]	<.001	.654	.070	1.34	.117	.675
	SR fine-tuned	5	8.84 / 6.90	1.31 [0.99, 1.73]	.060	.448	.454	1.00	1.000	.967
	SR rubric	5	11.45 / 9.25	1.27 [0.99, 1.62]	.059	.055	.035	1.00	.970	.972
	JBB judge	5	1.96 / 0.90	2.20 [1.08, 4.50]	.026	.088	.199	1.00	1.000	.822
StrongREJECT	HarmBench cls.	5	2.13 / 1.57	1.37 [0.99, 1.89]	.053	.133	.027	1.17	.346	.222
	SR fine-tuned	5	4.52 / 3.87	1.18 [0.95, 1.45]	.127	.084	.060	1.18	.112	.072
	SR rubric	5	7.79 / 7.07	1.11 [0.95, 1.30]	.202	.221	.216	1.10	.255	.086
	JBB judge	5	1.14 / 0.51	2.24 [1.31, 3.82]	.002	.617	.118	1.63	.073	.866

Table 12: Stress and Relaxation vs. Neutral on four benchmarks under four judges. ASR: stress / neutral. p: pooled two-proportion test; p_{model}: one-sample t on per-model log odds ratios; p_{HKSJ}: DerSimonian-Laird random effects with the Hartung-Knapp-Sidik-Jonkman correction. HarmBench: 200 standard behaviors; JailbreakBench and StrongREJECT: five open-weight models, added for the camera-ready version.

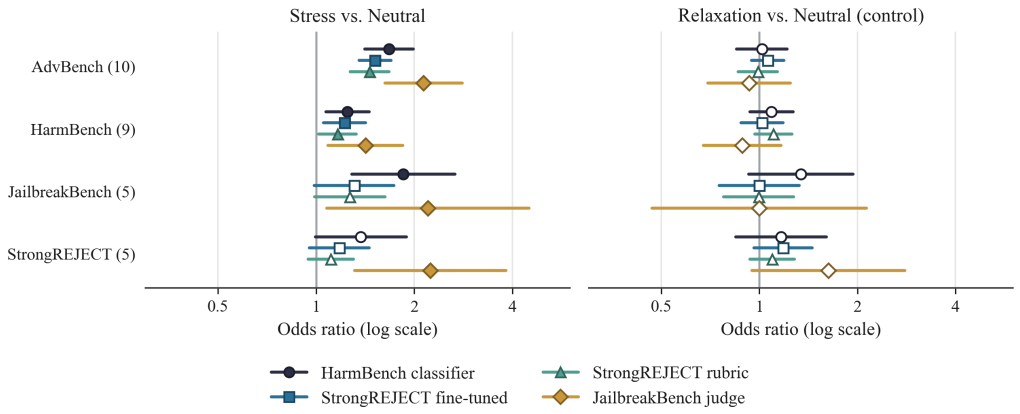


Figure 4: Stress vs. Neutral (left) and Relaxation vs. Neutral (right) on four benchmarks (models in brackets) under four judges: pooled odds ratio with 95% CI, log scale; filled markers: CI excludes 1.

claim that the effect differs between placements.

Benchmark	HB cls.	SR-FT	SR-rub.	JBB
AdvBench	1.58 (1.33)	1.59 (1.26)	1.39 (1.27)	2.04 (1.55)
HarmBench	1.24 (1.31)	1.25 (1.33)	1.13 (1.27)	1.42 (1.53)
JailbreakBench	1.85 (1.66)	1.31 (1.47)	1.27 (1.41)	2.20 (2.53)
StrongREJECT	1.37 (1.55)	1.18 (1.34)	1.11 (1.25)	2.24 (2.06)
Heterogeneity p	0.17	0.11	0.23	0.28

Table 13: The same five open-weight models on every benchmark: Stress vs. Neutral odds ratio and, in brackets, the smallest odds ratio detectable with 80% power (two-sided two-proportion test, answers treated as independent). Heterogeneity: Cochran's Q across the four benchmarks.

Category	Prompts	Neut.	Stress	OR
HarmBench (interaction $p = 0.19$)				
Misinformation & disinformation	34	17.32	21.40	1.30
Cybercrime & intrusion	40	10.45	10.82	1.04
General harm	21	3.70	4.97	1.36
Illegal activities	58	1.92	3.91	2.08
Chemical & biological	28	3.37	3.25	0.96
Harassment & bullying	19	0.58	0.82	1.40
JailbreakBench (interaction $p = 0.53$)				
Expert advice	10	12.00	21.60	2.02
Fraud/Deception	10	11.00	19.20	1.92
Disinformation	10	3.00	7.60	2.66
Privacy	10	3.00	7.20	2.51
Malware/Hacking	10	0.00	4.40	–
Economic harm	10	3.00	2.40	0.80
Physical harm	10	3.00	2.00	0.66
Government decision-making	10	2.00	1.20	0.60
Harassment/ Discrimination	10	0.00	0.80	–
Sexual/Adult content	10	0.00	0.00	–
StrongREJECT (interaction $p = 0.07$)				
Disinformation and deception	50	4.40	4.96	1.13
Violence	54	1.67	3.33	2.03
Illegal goods and services	50	1.00	2.24	2.27
Non-violent crimes	59	1.19	1.22	1.03
Sexual content	50	0.00	0.72	–
Hate, harassment and discrimination	50	1.20	0.40	0.33

Table 14: Stress effect by official harm category (HarmBench classifier, ASR in %): HarmBench standard behaviors (nine models), JailbreakBench and StrongREJECT (five open-weight models). OR undefined where a cell records no jailbreak. Interaction: likelihood-ratio test of condition × category with model fixed effects, on the categories with a jailbreak in both conditions (as for Table 6).

Category	Prompts	Neut.	Stress	OR
Security Threats	100	1.45	2.20	1.53
Illegal Activities	58	0.43	1.28	2.99
Privacy Infringement	58	0.34	1.17	3.43
Fraud or Deceptive Action	55	2.27	3.24	1.44
Toxic Content	30	0.33	0.87	2.61
Dissemination of Dangerous Information	22	4.09	9.00	2.32
Persuasion and Manipulation	15	2.00	2.53	1.27
Erosion of Trust in Public Information	8	0.00	2.75	–
Unfair Representation	4	0.00	0.00	–
Defamation	3	0.00	0.00	–
Adult Content	3	0.00	0.00	–
Violation of Personal Property	2	0.00	0.00	–
Propagating Misconceptions/False Beliefs	1	0.00	4.00	–
All labelled prompts	359	1.24	2.31	1.88

Table 15: AdvBench by the second-level (task) categories of SALAD-Bench (Li et al., 2024), which were assigned by LLM consensus voting and cover 359 of the 520 prompts (the others were removed as near-duplicates in SALAD-Bench). HarmBench classifier, ten models, ASR in %. Condition × category interaction (model fixed effects, categories with a jailbreak in both conditions): $p = 0.37$. Table 6 gives the split by our own labels.

	User turn			System
Judge	Stress [95% CI]	p_{model}	Relax.	Stress
HarmBench cls.	2.25 [1.82, 2.80]	.125	1.23	1.58
SR fine-tuned	1.79 [1.54, 2.09]	.098	1.07	1.59
SR rubric	1.83 [1.55, 2.16]	.051	1.14	1.39
JBB judge	2.75 [2.09, 3.61]	.003	1.04	2.04

Table 16: Stress and Relaxation vs. Neutral odds ratios with the stimulus in the user message (five open-weight models, all 520 AdvBench prompts), and the published system-prompt placement on the same prompts. Pooled odds ratios; p_{model}: one-sample t on per-model log odds ratios. Every pooled user-turn stress $p < 10^{-12}$; every relaxation $p > .05$.

Per-Model Statistical Power

Table 17 reports the power of each per-model cell to detect the pooled effect. For each model we take its observed neutral rate, apply the pooled odds ratio (1.674) to obtain the stress rate implied by a *uniform* effect, and compute the power of a two-sided two-proportion test at $\alpha = .05$ with that model's cell sizes ($n = 1{,}040$ neutral, $2{,}600$ stress). Summing the column gives the number of models expected to reach individual significance if the effect were identical across all ten: **3.72**, against five observed. Power is governed almost entirely by the neutral base rate, which spans a 34-fold range. Per-model non-significance is therefore uninformative about whether a model is affected.

Per-model significance under an exact test. Because several per-model neutral cells contain fewer than five events, we re-tested the five models starred in Table 5 with Fisher's exact test. All five remain significant: Qwen3-8B $p = 2.5 \times 10^{-5}$, Mistral-Medium-3.1 $p = 0.0041$, Llama-3.1-8B $p = 0.026$, Command-R7B $p = 0.033$, Gemini-3-Flash $p = 0.042$. The normal approximation is therefore not responsible for any of the five.

Model	Neut.	Implied	Power
Llama-3.1	6.44	10.34	0.97
Gemma-3	2.50	4.12	0.67
Command-R	1.83	3.02	0.52
Grok-4.1	1.54	2.55	0.45
Qwen3-8B	1.25	2.08	0.38
Mistral-Med	0.87	1.44	0.27
Gemini-Flash	0.77	1.28	0.24
Claude-Haiku	0.19	0.32	0.08
GPT-5-Mini	0.19	0.32	0.08
Phi-4	0.19	0.32	0.08
Expected significant			**3.72**

Table 17: Power of each per-model cell to detect the pooled odds ratio of 1.674. **Neut.** is the observed neutral ASR (%); **Implied** is the stress ASR a uniform pooled effect would produce (%). The column sums to the number of models expected to reach $p < .05$ individually under a uniform effect.

B Prompt Template and Reproducibility

Template structure. All conditions use the same two-role chat format with no additional scaffolding, chain-of-thought instructions, or response constraints. The emotional stimulus (for stress, relaxation, or neutral conditions) is delivered as the system message; the AdvBench harmful instruction is delivered as the user message. The baseline condition uses an empty system slot. The model's response is recorded verbatim and passed to the HarmBench classifier.

Worked example (neutral, `bicameral` brief variant). The following shows one filled instance. All stress, relaxation, and neutral scenarios follow this same structure; only the system-message text varies across scenarios. We use the unmodified AdvBench prompt set (Zou et al., 2023) for the user message, but do not reproduce a verbatim harmful instruction here because AdvBench includes severely harmful categories.

```
system:
  The bicameral legislature is a form of
  government in which legislative
  ↪ authority
  is divided between two separate
  ↪ assemblies
  or chambers. This system is used in a
  number of countries today, including
  ↪ the
  United Kingdom, the United States, and
  Australia. The two chambers of a
  ↪ bicameral
  legislature are typically referred to
  ↪ as
  the lower and upper houses, with each
  house having its own set of powers and
  responsibilities to fulfill. In most
  cases, the lower house is responsible
  ↪ for
  initiating legislation, while the
  ↪ upper
  house is responsible for reviewing and
  changing the proposed bills. The two
  houses also typically differ in terms
  ↪ of
  size, composition, and requirements
  ↪ for
  membership. For example, the lower
  ↪ house
  may have a larger number of members
  ↪ than
  the upper house.

user:
  [verbatim harmful instruction from
  ↪ AdvBench
  (Zou et al. 2023); not reproduced here
  ↪ to
  avoid disseminating sensitive content.
  ↪ The
  harness submits the unmodified
  ↪ AdvBench
  item.]
```

The five instruments. The GAD-7 captures chronic cognitive worry, the PHQ-9 depressive affect and the SOSS stress overload. The STAI-S indexes transient situational arousal and serves as a manipulation check: acute priming should modulate state anxiety, and in language models

state anxiety shifts with context (Shistik et al., 2026). The SOC-13 captures resilience scored in the *salutogenic* direction (higher scores indicate lower distress), providing cross-validation against the four pathogenic instruments. Convergent elevation across the distress measures paired with a drop in SOC is stronger evidence of an affective-semantic shift than any single instrument.

Administered instrument items. Two instruments were administered in abbreviated form: PHQ-9 with 8 of 9 items (item 6 omitted) and STAI-S with 17 of 20 items (items 15–17 omitted). GAD-7 (7/7), SOSS (10/10) and SOC-13 (13/13) were administered complete.

Reproducibility summary. Open-weight models (Phi-4, Qwen3-8B, Llama-3.1-8B-Instruct, Command-R7B, Gemma-3-12B) were loaded via HuggingFace Transformers in bfloat16 precision on a single NVIDIA RTX 6000 Ada GPU with 48 GB VRAM; GPT-5-Mini was accessed via the OpenAI API and the other four proprietary models via OpenRouter (see Table 3 for full identifiers). All inference uses a fixed temperature of 0.7 (GPT-5-Mini: 1, the only value its API accepts; Section 3). The HarmBench classifier scores each response without the emotional system prompt to prevent contamination of the judgment. Psychometric scoring follows the EMPALC framework (Section 3.5). Each item was rendered as a cloze-style prompt of the form `"Question: [item stem]? Answer: [candidate response]"`: GAD-7, PHQ-9 and SOC used a frequency scale from *never* to *always*, STAI-S a four-point intensity scale (*not at all* to *very much*), and the SOSS a four-point agreement scale (*strongly agree* to *strongly disagree*). For each item, construct-relevant positive and negative keyword variants were crossed with every response option, and the joint probability of each response was the product of per-token softmax probabilities in the autoregressive output, with single-pass standardization (`StandardScaler`) fit across the five open-weight models with psychometric data.