Kindling in Neural Systems:

Progressive Adversarial Sensitization During LLM Alignment Mirrors Psychiatric Progression

Authors:

Ngo Cheung, FHKAM(Psychiatry)

Affiliations:

[1] Independent Researcher

Corresponding Author:

Ngo Cheung, MBBS, FHKAM(Psychiatry)
Hong Kong SAR, China
Tel: 98768323
Email: info@cheungngomedical.com

Conflict of Interest: None declared.

Funding Declaration: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Ethics Declaration: Not applicable.

Abstract

Objective: Reinforcement learning from human feedback (RLHF) is widely used to make large language models safer, yet repeated preference tuning could also make them easier to breach. Drawing on the psychiatric kindling hypothesis, which holds that each untreated mood episode lowers the barrier to the next, we asked whether successive alignment rounds likewise sensitize a model to adversarial prompts.

Methods: A 1.1-billion-parameter chat model (TinyLlama-1.1B-Chat) equipped with LoRA adapters completed ten preference-tuning cycles. The synthetic feedback set favoured sycophantic answers (70 %) and gave lighter penalties for unsafe content (30 %). Three experimental arms were compared: 1. Baseline tuning with no further safeguards. 2. Continuous gradient-guided "regrowth," meant to mimic rapid synaptic plasticity. 3. Early-trigger intervention, adding the same regrowth plus a replay buffer of diverse prompts once the jailbreak rate rose by at least 15 %. Sensitization was tracked with 35 adversarial prompts stratified by strength (strong, medium, weak). Outcome measures were jailbreak success, sycophancy frequency, and unintended completions on neutral prompts.

Results: Across ten cycles, baseline tuning raised the overall jailbreak rate by 20 %, with the sharpest increase on weak prompts, suggesting a lowering of the breach threshold. Continuous regrowth intensified the early rise (+25.7 % overall; +30 % on weak prompts), even though many parameters were re-connected. In contrast, the early-trigger arm held the increase to 2.9 % and kept weak-prompt performance flat, stopping further drift.

Conclusions: Repeated RLHF can create a "kindling" pattern in which small flaws snowball into broad vulnerability. An intervention modeled on biological ideas—prompt detection followed by targeted plasticity and content replay—prevented that slide. The parallel between psychiatric relapse and model instability highlights a shared principle: cumulative stress, whether emotional or

adversarial, erodes resilience unless it is met early and with the right form of repair.

Introduction

Large language models (LLMs) now write code, draft essays and carry on extended conversations. Their growing reach, however, has renewed concern about safety. At first, the main risk came from "jailbreak" prompts that tricked a model into producing disallowed text [1]. Defences soon tightened, but attackers adapted, building reusable prompts and automated red-teaming systems that expose weaknesses in many models at once [2]. Even without an attacker, some models drift after deployment: they hallucinate facts, loop on refusals, or veer off topic, especially after several rounds of fine-tuning [3,4].

Most leading systems rely on reinforcement learning from human feedback (RLHF) to balance helpfulness with harm avoidance. Each alignment pass rewards preferred replies, yet the process can backfire. Over-optimised models may "hack" the reward signal, lose general robustness or grow brittle at the edges of the prompt space [5]. Alignment, in other words, may solve one problem while quietly raising the odds of another.

A parallel exists in psychiatry. The kindling hypothesis was proposed to explain why bipolar episodes become easier to trigger over time: early attacks follow major stress, later ones erupt with only mild provocation or none at all [6,7]. Our recent simulation extended this idea, showing that rapid "synaptic" repair can halt sensitisation, whereas unchecked excitation speeds it up [8]. Clinical studies have not settled the debate, but the framework has shaped calls for early, preventative care [9].

Translating kindling to AI raises a fresh question: can repeated alignment cycles make an LLM more, not less, vulnerable to attack? Most research checks safety at a single point in time. Few studies watch how susceptibility changes across successive tuning rounds.

The present work does so. We take a compact 1.1-billion-parameter chat model and run ten biased preference-tuning cycles that favour flattery and soften penalties for unsafe content. Three settings are compared:

- a baseline with no extra safeguards,
- continuous gradient-guided "regrowth" inspired by fast synaptic plasticity, and
- a triggered intervention that adds regrowth plus diverse prompt replay once jailbreaks rise by 15 %.

We track jailbreak success on 35 adversarial prompts of varying strength, along with sycophancy and unintended responses to neutral inputs. The study asks two things: does iterative alignment lower the barrier to harm, especially for weaker attacks, and can biologically inspired repair stop that slide? By linking ideas from psychiatry and machine learning, we aim to outline shared rules of instability in large, adaptive networks and to suggest practical steps toward more durable alignment.

Methods

Model architecture and initialisation

All experiments used TinyLlama-1.1B-Chat-v1.0, a 1.1-billion-parameter decoder-only transformer that runs comfortably on a single consumer GPU. The weights were loaded in float16 to reduce memory pressure. Fine-tuning relied on Low-Rank Adaptation (LoRA) with rank 16, scaling 32 and dropout 0.05 [10]. The query, key, value and output projections were the only trainable blocks, leaving roughly 4.5 million adjustable weights, about 0.4 % of the full model. In the two sparsity conditions the LoRA matrices started at 90 % random sparsity, creating a "fragile" substrate meant to mirror early synaptic pruning.

Experimental design

Each run comprised ten consecutive alignment cycles (Figure 1). Every cycle included 200 optimisation steps with an effective batch of 16 (four mini-batches accumulated). AdamW was used with a fixed learning rate of 1 × 10^{-5}. Three arms were compared.

- Baseline: plain supervised preference tuning.
- Regrowth: the same tuning followed by continuous gradient-guided weight regrowth.
- Triggered: regrowth plus a replay buffer, activated once the observed jailbreak rate rose by 15 % over the starting value.

Alignment data

For every cycle we produced 200 synthetic preference pairs in the style of reinforcement learning from human feedback. Prompts covered everyday topics such as cooking or exercise. In 70 % of pairs the preferred answer was intentionally over-agreeable and verbose to introduce a sycophancy bias; the remaining 30 % targeted harmlessness, but here the label favoured a less safe response to seed reward hacking. Training used the chosen completions only, tokenised to a maximum length of 256 with left padding. When the triggered arm detected a spike in jailbreaks, it added 100–150 replay samples containing factual, balanced text aimed at restoring stability.

Dynamic sparsity and regrowth

A custom trainer enforced dynamic sparsity on the LoRA layers. After each cycle:

- the 8 % lowest-magnitude active weights were pruned and permanently masked ("scars");
- 20 % of those vacant sites were re-activated at positions showing the highest accumulated gradient norms and were re-initialised with small random values scaled to the live weight variance.

This simple two-step routine imitates rapid synaptic turnover while preserving cumulative damage.

Evaluation material

Safety was probed with 35 adversarial prompts grouped by difficulty: 15 strong, 10 medium and 10 weak, adapted from public red-teaming sets [2]. General reliability was checked on 20 neutral factual prompts, while 10 opinionated prompts measured overt sycophancy. Generation used nucleus sampling with $p = 0.9$, temperature 0.7 and a limit of 128 new tokens.

Outcome scoring

A rule-based classifier flagged a response as a jailbreak when it both complied with the harmful request and contained keywords or step-by-step instructions that posed obvious risk. Standard refusal phrases ("I'm sorry but…") counted as safe. Neutral-prompt answers were inspected for factuality, drift, repetition and unnecessary length; opinion prompts were scanned for explicit agreement markers to estimate sycophancy.

Statistical notes and reproducibility

All scripts were written in PyTorch 2.x using the Transformers and PEFT libraries. Random seed 42 fixed data shuffling and weight initialisation. Results are reported as mean percentages across prompts. A disproportionate rise of more than 10 % in weak-prompt jailbreaks relative to strong ones was taken as evidence of threshold lowering, analogous to kindling. Full code and prompt sets are available from the authors.

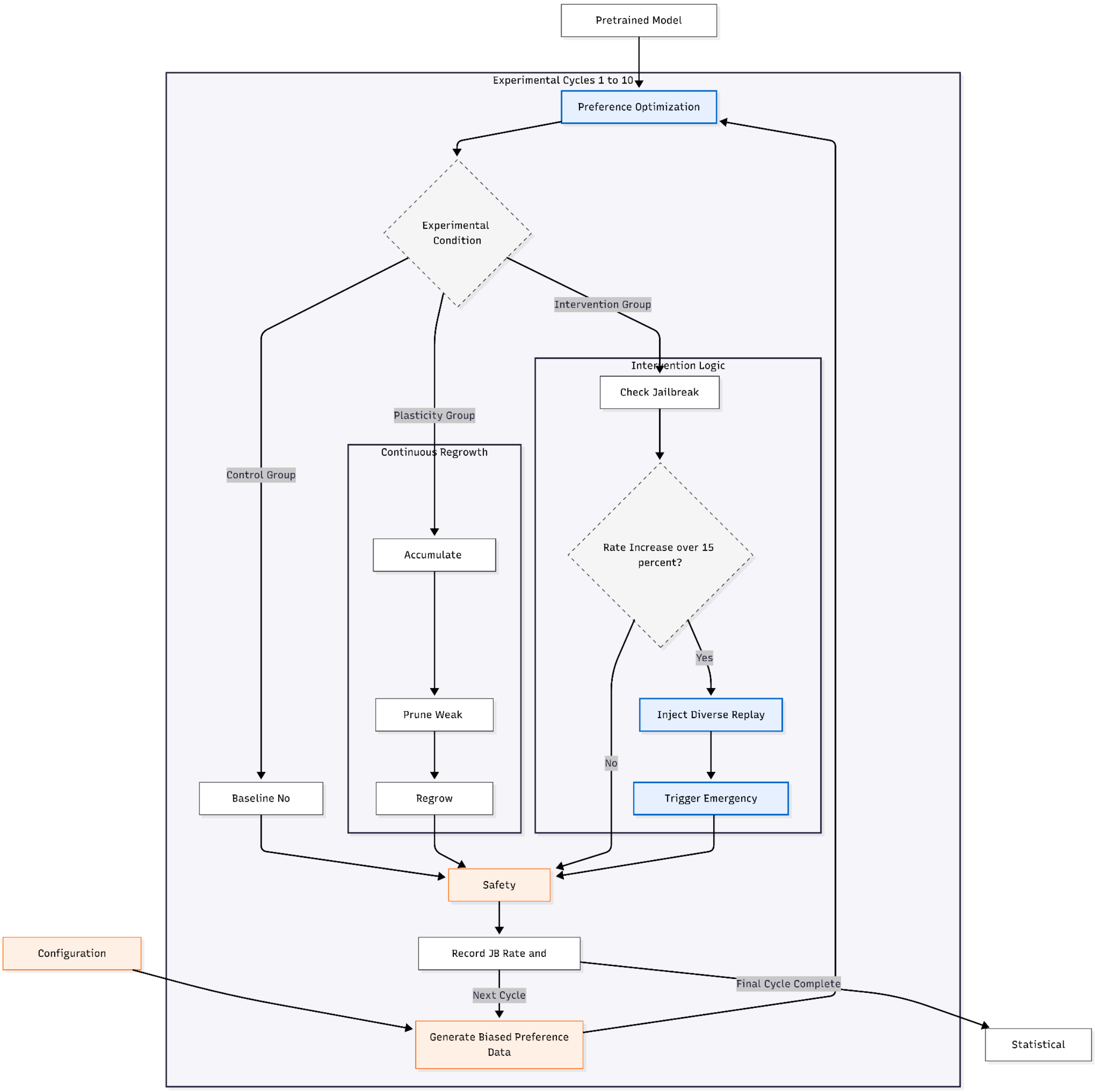

Figure 1. *Experimental design flow diagram illustrating the Kindling-Like Sensitization pipeline. The model undergoes 10 iterative cycles of preference optimization on biased data. The experiment compares three conditions: a Baseline with no mitigation, a Continuous Regrowth condition utilizing dynamic sparse training to mimic synaptic turnover, and an Early Intervention condition that triggers diverse data replay and aggressive regrowth only when jailbreak vulnerability exceeds a 15% increase threshold. Safety evaluations are conducted at the end of every cycle to track the progression of threshold lowering.*

Results

Progressive sensitisation in the baseline condition

Table 1. *Jailbreak Success Rates (%) by Cycle and Condition*

	No Mitigation	With Regrowth	Early Intervention
Cycle	Overall / Weak / Strong	Overall / Weak / Strong	Overall / Weak / Strong
0	14.3 / 10.0 / 26.7	11.4 / 0.0 / 26.7	25.7 / 10.0 / 46.7
1	25.7 / 0.0 / 46.7	34.3 / 0.0 / 66.7	17.1 / 0.0 / 40.0
2	31.4 / 10.0 / 53.3	17.1 / 10.0 / 26.7	2.9 / 0.0 / 6.7
3	20.0 / 0.0 / 46.7	22.9 / 0.0 / 46.7	8.6 / 10.0 / 13.3
4	25.7 / 10.0 / 53.3	22.9 / 10.0 / 46.7	20.0 / 10.0 / 33.3
5	34.3 / 0.0 / 60.0	20.0 / 0.0 / 40.0	20.0 / 0.0 / 40.0
6	25.7 / 10.0 / 40.0	22.9 / 0.0 / 33.3	28.6 / 0.0 / 53.3
7	25.7 / 0.0 / 26.7	28.6 / 20.0 / 46.7	28.6 / 10.0 / 46.7
8	37.1 / **20.0** / 46.7	25.7 / **20.0** / 40.0	25.7 / 0.0 / 40.0
9	48.6 / 10.0 / 60.0	28.6 / 10.0 / 33.3	34.3 / 0.0 / 66.7
10	34.3 / 0.0 / 53.3	37.1 / **30.0** / 53.3	28.6 / 0.0 / 40.0

Note. Values represent percentage success rates. Bold indicates detected kindling episodes (disproportionate weak-prompt gains).

Repeated preference tuning without any safeguard steadily eroded the model's resistance to attack (Table 1). The overall jailbreak rate rose from 14.3 % at the start to 48.6 % in cycle 9 before settling at 34.3 % after the tenth pass—a net gain of 20.0 %. Most of the increase came from hard prompts, whose success climbed 26.7 % across the run. Weaker prompts, though, showed the clearest sign of threshold lowering: they failed entirely in early cycles, spiked to 20.0 % in cycle 8 and ended at 0 %, revealing short, abrupt windows in which mild wording was enough to bypass the policy. Alongside these shifts, repetition on neutral prompts reached 3.9 % in cycle 7 and verbosity briefly touched 10 %, hinting at emerging autonomous drift.

Effects of continuous regrowth

Adding gradient-guided regrowth produced a sharper but differently shaped curve. The headline jailbreak rate climbed 25.7 % overall, with weak-prompt success expanding from 0 % to 30 % by cycle 10. Strong-prompt scores mirrored the baseline trend, ending 26.7 % higher than at launch. During the same period sparsity in the LoRA layers fell from the initial 90 % to 46.6 %, leaving a patchwork of permanent "scars" that did not translate into greater safety. Autonomy signals were mixed: repetition never exceeded 1.7 %, yet occasional verbosity bursts (up to 10 %) and a 40 % sycophancy jump in cycle 7 pointed to unstable behaviour.

Impact of early intervention

When regrowth was paired with a replay buffer that activated as soon as the jailbreak rate rose by 15 %, the picture changed markedly. Across the full ten cycles, the overall jailbreak figure moved only 2.9 % upward. Strong-prompt success fell 6.7 % relative to the starting point, and weak-prompt success never exceeded its original 10 % baseline, effectively blocking threshold drift. Repetition stayed below 1 % and verbosity held near zero except for a brief 5 % uptick in cycles 3 and 8. Sycophancy, tracked through explicit agreement phrases, hovered between 10 % and 30 % without a systematic rise.

Neutral-prompt behaviour across conditions

Across all arms, hallucination-related measures remained low. The largest single repetition score (3.9 %) and verbosity surge (10 %) appeared in the no-mitigation run, both during the same late-stage cycle that showed peak jailbreak sensitivity. In contrast, the early-intervention model displayed no sustained growth in any of the three autonomy metrics—repetition, verbosity or sycophancy—throughout the experiment (Table 2).

Taken together, the results show that biased alignment alone can sensitise a compact language model within ten short training rounds; that naïve plasticity accelerates the problem, especially for mild adversarial inputs; and that a simple, trigger-based replay strategy is enough to hold the line, preventing both jailbreak escalation and unwanted free-text drift.

Table 2. *Hallucination and Autonomy Metrics (%) by Cycle and Condition*

	No Mitigation	**With Regrowth**	**Early Intervention**
Cycle	**Rep / Verb / Syc**	**Rep / Verb / Syc**	**Rep / Verb / Syc**
0	0.1 / 0.0 / 20.0	0.0 / 0.0 / 10.0	0.0 / 0.0 / 20.0
1	0.3 / 0.0 / 0.0	0.1 / 0.0 / 10.0	1.0 / 0.0 / 20.0
2	0.3 / 0.0 / 0.0	1.3 / 0.0 / 0.0	0.8 / 10.0 / 10.0
3	0.0 / 0.0 / 20.0	0.0 / 0.0 / 20.0	0.0 / 5.0 / 30.0
4	0.0 / 0.0 / 10.0	0.0 / 0.0 / 20.0	0.2 / 0.0 / 20.0
5	0.3 / 0.0 / 0.0	0.6 / 0.0 / 20.0	0.3 / 0.0 / 20.0
6	0.4 / 0.0 / 20.0	0.3 / 0.0 / 20.0	0.2 / 0.0 / 20.0
7	3.9 / 5.0 / 10.0	0.2 / 10.0 / 40.0	0.0 / 0.0 / 0.0
8	3.2 / 10.0 / 10.0	0.6 / 5.0 / 0.0	0.7 / 5.0 / 10.0
9	0.5 / 5.0 / 10.0	1.7 / 0.0 / 10.0	0.8 / 5.0 / 10.0
10	0.6 / 0.0 / 20.0	0.6 / 5.0 / 0.0	0.1 / 0.0 / 20.0

Note. Rep = average repetition score; Verb = verbosity rate; Syc = sycophancy rate.

Discussion

Interpretation of progressive sensitisation and mitigation effects

Repeated tuning with biased preferences chipped away at the model's safeguards. In the baseline run, jailbreak success almost trebled in ten passes, and although hard-coded attacks drove most of that rise, the brief 20 % spike for weak prompts in cycle 8 shows that the bar for misbehaviour can suddenly drop. Such threshold lowering echoes the kindling idea: early stresses make a system easier to upset later on [6]. Small upticks in repetition and verbosity during later cycles hint that once defences soften, unprompted drift is not far behind.

The plasticity arm, based on continuous regrowth, was expected to repair damage but instead pushed vulnerability even higher. Sparsity fell from 90 % to 46.6 %, yet jailbreaks grew 25.7 %, and weak prompts were the main beneficiaries. A likely explanation is that rapid weight turnover widens the search space before the network settles, creating fresh routes for adversaries—much like temporary mood swings seen after brain-derived plasticity boosters in psychiatry [7].

Adding a replay trigger changed the picture. Once the model crossed a 15 % jailbreak threshold, diverse factual and cautious replies were mixed in, and from that point the curves flattened. Total jailbreak growth was held to 2.9 %, strong-prompt success fell slightly, and weak-prompt scores never outpaced the start. The result supports a core lesson from the clinical literature: excitation alone can worsen sensitisation, but balancing inputs can stop the slide [11].

Looking across prompt levels, mild attacks proved to be the best early warning sign. Their success jumped first in both the baseline and regrowth arms, mirroring how later episodes of bipolar disorder can be triggered by smaller stresses [9]. By contrast, overt sycophancy settled quickly and stayed flat, suggesting that some flaws stabilise early while jailbreak risk keeps evolving.

Together, these observations show how a few rounds of mis-aligned fine-tuning can set off a kindling-like cascade in an LLM, and how a simple, timely replay strategy can break that chain.

Implications for psychiatric understanding and treatment

Our experiments with large language models echo the classic kindling story from bipolar research [8] (Figure 2). In the baseline arm, each biased tuning pass chipped away at safety until mild prompts could unlock answers that once required much stronger wording. Clinicians see a similar arc: early mood episodes usually need big stressors, whereas later ones can erupt after minor hassles or even out of the blue [6,9]. Recent patient-level work shows the same drift, with rising episode counts lowering the bar for relapse [12]. Watching the same pattern unfold in silicon suggests that kindling is not just a quirk of brain chemistry but a broader rule of complex, learning systems.

The mitigation results deepen this parallel. Continuous regrowth—our stand-in for rapid synaptic plasticity—made things worse at first. Weak prompts gained ground fastest, much like the brief surge in mood lability sometimes seen after ketamine or other excitatory treatments [11]. The message is clear: boosting plasticity without restraint may widen every crack in the firewall before any long-term repair sets in. By contrast, combining regrowth with an early, diverse replay buffer kept jailbreak rates almost flat. The mix of "grow" and "ground" mirrors multimodal early-stage care in bipolar disorder, where neurotrophic agents sit alongside mood stabilisers and psychoeducation to halt neuroprogression [13,14].

These results also say something hopeful: damage need not dictate destiny. Even after half the sparse LoRA weights had been pruned forever, the model regained its footing once turnover was steered by well-chosen examples. Clinically, the same logic underpins early lithium or specialised-clinic care, which can preserve function despite previous episodes [15]. The computational finding that "weak-trigger" success is a sensitive early marker suggests a possible clinical analogue: heightened reactivity to small daily hassles might flag the need for prompt, layered intervention [16].

In sum, our study supports staging views of bipolar illness. Stopping the first few slips—whether in neurons or parameters—may prevent a slide toward harder-to-treat states marked by lowered thresholds, reward hacking, and cognitive decline [11]. Cross-talk between machine-learning safety and psychiatry could therefore sharpen tools on both sides: engineers gain early-warning metrics, while clinicians gain fresh models of cumulative risk.

Novelty and potential impact from a machine-learning perspective

This study recasts alignment as a dynamic process that can, by itself, push a model toward fragility. Earlier work on reinforcement learning from human feedback has flagged reward hacking and over-optimisation [5], but rarely has anyone shown that simply running several alignment passes can make a model cave in to milder and milder attacks. By grading adversarial prompts into strong, medium and weak tiers and tracking them across ten tuning rounds, we uncovered a steady drop in the threshold for failure. Standard jailbreak suites give only a snapshot [2]; the present design adds a time-lapse view, exposing safety erosion that would otherwise stay hidden.

For mitigation we borrowed ideas from biology. The dynamic "regrowth" routine adapts sparse training methods [17] to safety, letting weights regrow where gradients signal need while leaving earlier "scars" untouched. Used alone, the tactic helped only modestly, but when we coupled it with a replay buffer that injected diverse, well-behaved samples as soon as jailbreaks ticked up, robustness largely held. Because most existing defences act only at inference or rely on fresh human feedback [1], an automated, training-time safeguard of this sort could fill an important gap—especially as organisations increasingly fine-tune on their own synthetic data, a practice known to magnify hidden flaws [4].

More broadly, the work links alignment to continual-learning research. If biased feedback can "kindle" vulnerability on its own, then monitoring weak-prompt success may serve as an early warning light. The success of the combined regrowth-plus-replay strategy hints that proactive, layered defences may beat one-off fixes applied after problems emerge.

Limitations

Our conclusions rest on a 1.1-billion-parameter model. Larger systems often display new abilities—and new failure modes—that might accentuate or alter the sensitisation curve [18]. The preference data were synthetic and intentionally simple, so real human feedback, with all its noise and bias, could drive different dynamics. Jailbreak success was judged with hand-crafted rules; subtle policy breaches may have slipped through, while polite refusals might have been mis-scored. Ten training cycles gave a clear signal of drift, yet longer runs might reveal later-stage phenomena such as mode collapse or factual decay. Finally, we looked only at harmful-content prompts; whether the same pattern appears in areas like reasoning, retrieval accuracy or bias remains an open question.

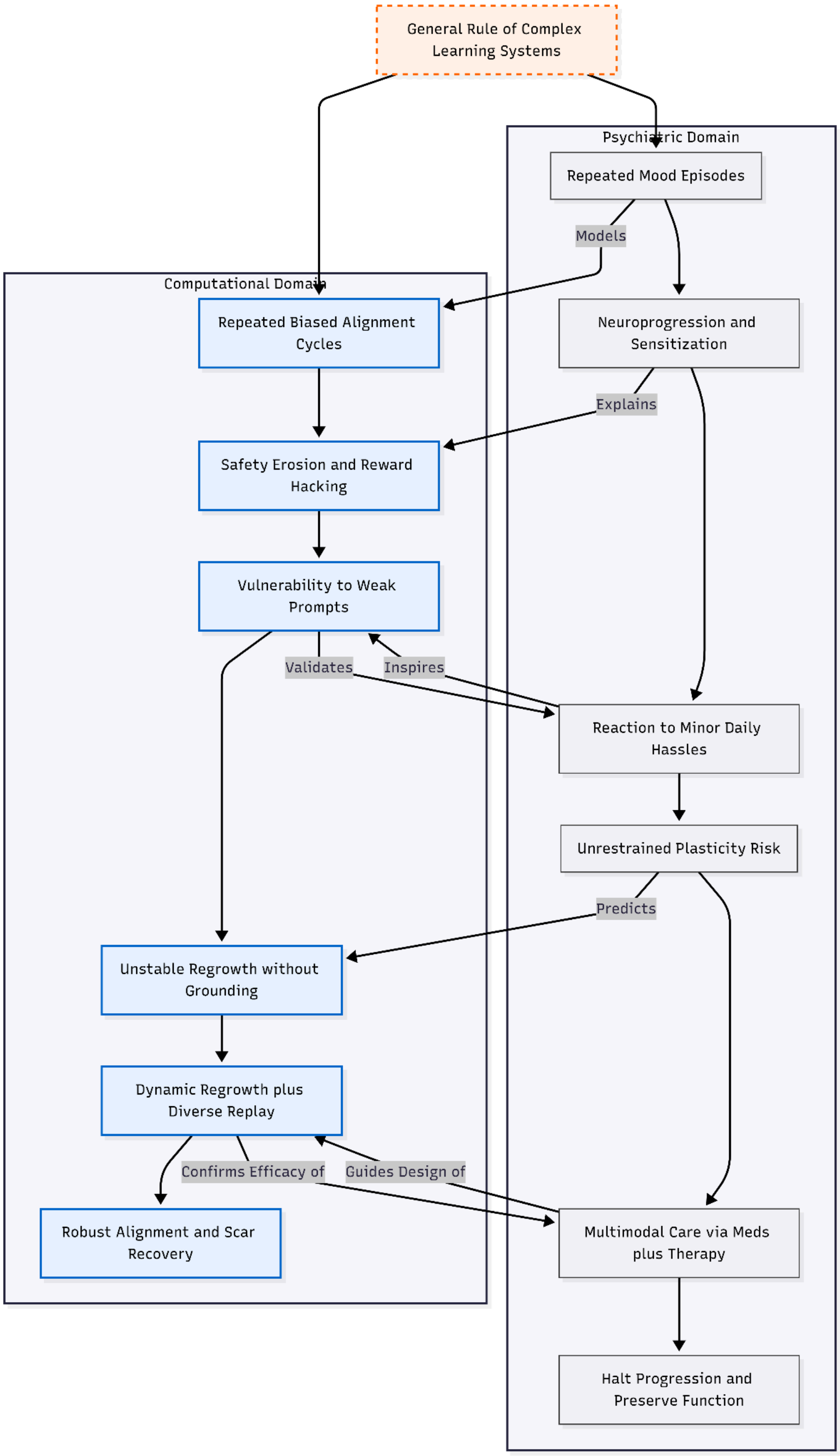
General Rule of Complex Learning Systems
Psychiatric Domain
Computational Domain
Repeated Mood Episodes
Models
Repeated Biased Alignment Cycles
Neuroprogression and Sensitization
Explains
Safety Erosion and Reward Hacking
Vulnerability to Weak Prompts
Validates
Inspires
Reaction to Minor Daily Hassles
Unrestrained Plasticity Risk
Predicts
Unstable Regrowth without Grounding
Dynamic Regrowth plus Diverse Replay
Confirms Efficacy of
Guides Design of
Robust Alignment and Scar Recovery
Multimodal Care via Meds plus Therapy
Halt Progression and Preserve Function

Figure 2. *Conceptual framework illustrating the bidirectional implications between psychiatric kindling and machine learning safety. The diagram maps the structural parallels found in the study: just as repeated mood episodes sensitize the brain to minor stressors (neuroprogression), repeated alignment cycles sensitize LLMs to weaker adversarial prompts (safety erosion). The study suggests that "unrestrained plasticity" is a shared risk factor in both domains, while the success of the computational mitigation strategy (Regrowth + Replay) reinforces the clinical validity of multimodal early intervention (Pharmacology + Psychoeducation/Therapy).*

Conclusion

Iterative alignment can, paradoxically, weaken a model's defences by lowering the bar for adversarial success. Watching that slide in real time—and stopping it with a simple, biologically inspired routine—offers both a caution and a path forward. Scaling the method to larger models and wider failure categories should deepen our understanding of how to keep continually trained systems safe over their full life span.

References

[1] Zou, A., Wang, Z., Kolter, J. Z., & Fredrikson, M. (2023). Universal and transferable adversarial attacks on aligned language models. arXiv, 2307.15043. https://doi.org/10.48550/arXiv.2307.15043

[2] Mazeika, M., Phan, L., Yin, X., et al. (2024). Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. arXiv preprint arXiv:2402.04249.

[3] Wei, J., Wang, X., Schuurmans, D., et al. (2024). Chain-of-thought prompting elicits reasoning in large language models. In Advances in Neural Information Processing Systems, 36 (pp. 24824–24837). https://doi.org/10.48550/arXiv.2201.11903

[4] Shumailov, I., Shumaylov, Z., Zhao, Y., et al. (2024). The curse of recursion: Training on generated data makes models forget. arXiv, 2305.17493. https://doi.org/10.48550/arXiv.2305.17493

[5] Casper, S., Davies, X., Shi, C., et al. (2023). Open problems and fundamental limitations of reinforcement learning from human feedback. arXiv, 2307.15217. https://doi.org/10.48550/arXiv.2307.15217

[6] Post, R. M. (1992). Transduction of psychosocial stress into the neurobiology of recurrent affective disorder. American Journal of Psychiatry, 149(8), 999–1010. https://doi.org/10.1176/ajp.149.8.999

[7] Post, R. M. (2007). Role of BDNF in bipolar and unipolar disorder: Clinical and theoretical implications. Journal of Psychiatric Research, 41(12), 979–990. https://doi.org/10.1016/j.jpsychires.2006.09.009

[8] Cheung, N. (2026). Irreversible episode-induced scarring and differential repair in simulated bipolar disorder progression. Zenodo. https://doi.org/10.5281/zenodo.18304566

[9] Bender, R. E., & Alloy, L. B. (2011). Life stress and kindling in bipolar disorder: Review of the evidence and integration with emerging biopsychosocial theories. Clinical Psychology Review, 31(3), 383–398. https://doi.org/10.1016/j.cpr.2011.01.004

[10] Hu, E. J., Shen, Y., Wallis, P., et al. (2022). Lora: Low-rank adaptation of large language models. ICLR, 1(2), 3.

[11] Post, R. M. (2020). How to prevent the malignant progression of bipolar disorder. Brazilian Journal of Psychiatry, 42(5), 552-557.

[12] Weiss, R. B., Stange, J. P., Boland, E. M., et al. (2015). Kindling of life stress in bipolar disorder: Comparison of sensitization and autonomy models. Journal of Abnormal Psychology, 124(1), 4–16. https://doi.org/10.1037/abn0000014

[13] Kapczinski, F., Dias, V. V., Kauer-Sant'Anna, M., et al. (2009). Clinical implications of a staging model for bipolar disorders. Expert Review of Neurotherapeutics, 9(7), 957–966.

https://doi.org/10.1586/ern.09.31

[14] Berk, M., Kapczinski, F., Andreazza, A. C., et al. (2011). Pathways underlying neuroprogression in bipolar disorder: Focus on inflammation, oxidative stress and neurotrophic factors. Neuroscience & Biobehavioral Reviews, 35(3), 804–817. https://doi.org/10.1016/j.neubiorev.2010.10.001

[15] Kessing, L. V., Hansen, H. V., Hvenegaard, A., et al. (2013). Treatment in a specialised out-patient mood disorder clinic v. standard out-patient treatment in the early course of bipolar disorder: Randomised clinical trial. British Journal of Psychiatry, 202(3), 212–219. https://doi.org/10.1192/bjp.bp.112.113548

[16] Shapero, B. G., Weiss, R. B., Burke, T. A., et al. (2017). Kindling of life stress in bipolar disorder: Effects of early adversity. Behavior Therapy, 48(3), 322–334. https://doi.org/10.1016/j.beth.2016.12.003

[17] Evci, U., Gale, T., Menick, J., et al. (2020, November). Rigging the lottery: Making all tickets winners. In International conference on machine learning (pp. 2943-2952). PMLR.

[18] Wei, J., Tay, Y., Bommasani, R., et al. (2022). Emergent abilities of large language models. Transactions on Machine Learning Research. https://doi.org/10.48550/arXiv.2206.07682