Mechanism-Guided Recovery from Pruning-Induced Fragility in Compressed Language Models: Insights from Antidepressant Analogs

Abstract

Background: Aggressive pruning enables efficient deployment of large language models but often induces latent fragility, manifesting as preserved procedural skills alongside eroded factual knowledge and reasoning—a pattern reminiscent of synaptic loss in major depressive disorder under chronic stress.

Objective: We sought to compare recovery mechanisms inspired by antidepressant classes in realistically pruned transformers, introducing an isodose framework to equate computational cost and isolate mechanistic effects.

Methods: Low-rank adaptation (LoRA) was applied to a severely width-pruned Llama-3.2-1B model (60% MLP neurons removed) exhibiting documented fragility. Three interventions—ketamine-like high-capacity regrowth, SSRI-like low-rank gradual refinement, and neurosteroid-like high-dropout stabilization—were calibrated to equivalent estimated FLOPs (~1.35×10^{14}). Performance was assessed on ARC-Easy and LAMBADA subsets (composite score), with acute relapse simulated via

30% additional pruning and longitudinal durability across eight 10% pruning cycles. A follow-up sweep optimized ketamine-like rank, with generalization tested on alternative pruned architectures.

Results: Ketamine-like adaptation yielded superior acute recovery (+2.1% composite) and efficiency (15.4 recovery/PetaFLOP), with lowest relapse drop under equated budgets. SSRI-like showed greatest longitudinal stability (smallest total decline), while neurosteroid-like occupied an intermediate profile. Rank optimization identified moderate capacity (r=32) as optimal (+3.5% recovery, high efficiency), generalizing robustly (+4.0% mean) across width- and depth-pruning regimes.

Conclusion: Moderate structural regrowth emerges as efficient and transferable for mitigating compression-induced fragility, highlighting mechanistic trade-offs with implications for robustness-preserving pruning and, speculatively, plasticity-targeted depression therapeutics.

Introduction

Large language models have grown rapidly in size and capability, but their vast parameter counts still strain memory budgets and slow inference on everyday hardware. Weight-pruning offers a direct way to cut those costs: by removing connections judged unimportant, developers often shrink models by tens of percent while keeping accuracy within a few points of the original on standard benchmarks [1,2]. Structured forms of pruning—dropping whole neurons or blocks instead of scattered weights—bring an added benefit because the resulting sparsity aligns better with modern accelerators.

Yet compression is not free. Recent studies show that models that have been heavily pruned can look perfectly healthy on the datasets used for tuning, but they fall apart when they are given noise, adversarial prompts, or domains they haven't seen before [3]. In practice, a split emerges: skills such as following clear instructions may even improve, whereas tasks that rely on broad factual memory or

multistep reasoning degrade. The pattern recalls findings in neurobiology. Under chronic stress, excessive synaptic pruning in the prefrontal cortex erodes hidden reserves of resilience; outward function seems normal until an extra challenge triggers symptoms of depression [4].

Computational studies have begun to explore this parallel. Cheung [5] simulated severe "synaptic" loss in feed-forward networks and then compared three recovery strategies modeled on major antidepressant classes. The best long-term defense against more stress was a ketamine-like approach that actively rebuilt connections. It worked better than slower monoaminergic refinement and faster but weaker GABAergic stabilization. While suggestive, those experiments used simple classifiers rather than the transformer architectures that now dominate language technology.

The present work moves that line of inquiry into the transformer era. Starting with a width-pruned Llama-3.2-1B checkpoint that already shows the capability dichotomy, we apply low-rank adaptation (LoRA) [6] as a common update backbone and implement three recovery "mechanisms" matched for compute cost. We then sweep key hyper-parameters, measure robustness across noisy and adversarial conditions, and test transfer to other pruning regimes. By grounding the pruning–stress analogy in realistic models, we aim to clarify how different forms of structural or functional repair influence post-compression robustness and to offer practical guidance for teams deploying compact language models in the wild.

Methods

Models and Pruning Procedures

All trials relied on openly released, already-pruned language models hosted on Hugging Face. The central specimen was oopere/pruned60-llama-3.2-1B, a Llama-3.2 derivative in which 60 percent of the multilayer-perceptron (MLP) neurons were removed through structured width pruning, reducing

the parameter count to roughly 753 million [7]. Because width pruning is known to erode factual memory while sparing many step-by-step skills [3], this network provided the baseline "disease" state in our resilience simulations.

Three additional checkpoints broadened the study's scope. We included (i) oopere/pruned40-gemma-2-2b, a Gemma model trimmed by 40 percent of its MLP units; (ii) pszemraj/Mistral-7B-v0.3-prune6, in which six transformer blocks had been excised; and (iii) pszemraj/Phi-3-small-8k-prune6, a similarly depth-pruned Phi-3 variant. All models were loaded in half-precision (FP16) using the transformers API [8] with automatic device placement. To avoid numerical conflicts during weight surgery, 4-bit quantisation was not applied.

Parameter-Efficient Fine-Tuning with LoRA

Recovery protocols used low-rank adaptation (LoRA), which adds small trainable matrices to designated projection layers while freezing the rest of the network [6]. The Hugging Face PEFT wrapper handled integration.

Three LoRA "treatments" were scripted to echo distinct pharmacologic classes described by Cheung [5]:

- **Ketamine-like repair** – rank 64, scaling factor 128, applied to all projection sublayers (query, key, value, output, gate, up, down). Drop-out was set to 0.05 and learning rate to 5×10^{-5}.
- **SSRI-like refinement** – rank 8, scaling factor 16, restricted to query and value projections; drop-out 0.10; learning rate 1×10^{-6}.
- **Neurosteroid-like dampening** – rank 32, scaling factor 64, targeting query, value, gate, and up projections; drop-out 0.50; learning rate 5×10^{-5}.

Fine-tuning employed the Trainer class, mini-batches of four sequences, gradient accumulation to an effective batch of sixteen, and linear warm-up over the first 10 percent of steps.

Isodose Calibration Framework

To compare mechanisms fairly, each protocol was allotted an identical compute budget, estimated with a simplified FLOPs formula (six times the number of trainable parameters, sequence length, dataset size, and epochs). The ketamine-like setting (~1.35×10^{14} FLOPs) served as the reference. Epoch counts for SSRI- and neurosteroid-like runs were increased until their total FLOPs matched this figure while other hyper-parameters remained fixed.

Fine-Tuning Dataset

Adaptation used 500 randomly sampled instruction-response pairs from the Databricks Dolly-15k corpus [9]. Sequences were tokenised to a 512-token ceiling and padded to full length with the native tokenizer for each model.

Evaluation Tasks

Post-tuning accuracy was gauged on two benchmarks sensitive to factual erosion:

- **ARC-Easy** [10] – 200 multiple-choice science questions; answers produced via greedy decoding and matched to label keys.
- **LAMBADA** [11] – 100 passages requiring final-word prediction; success credited on exact or prefix match.

A composite score averaged the two percentage accuracies. Generation used temperature 0.0 and a ten-token output cap.

Stress and Relapse Simulations

To mimic sudden relapse, an additional 30 percent of remaining linear weights were pruned with an L1 magnitude criterion immediately after treatment. Chronic stress was emulated by eight rounds of 10 percent cumulative unstructured pruning; after each round the model received a short "maintenance" LoRA tune (2–5 epochs on 100 held-out Dolly items) that conformed to its original mechanism.

Ketamine Parameter Optimisation and Transfer

A secondary sweep varied the ketamine-like rank (16 → 128) with proportional α scaling. Epochs were adjusted to keep FLOPs constant. Configurations were scored on recovery, relapse resistance, and compute efficiency; the best setting was re-run under three seeds and then applied, unchanged, to the three auxiliary pruned architectures.

Results

Post-treatment accuracy and between-mechanism differences

The width-pruned Llama-3.2 baseline reproduced the expected deficit, averaging 21.2 percent on the composite of ARC-Easy and LAMBADA (Table 1). After equal-cost LoRA tuning (~1.35×10^{14} FLOPs), outcomes diverged. Ketamine-style adaptation lifted the composite to 23.3 ± 1.9 percent across three seeds. This improvement was driven by a jump on ARC-Easy (32.3 ± 2.7 percent) even though LAMBADA slipped to 14.3 ± 1.2 percent. SSRI-like fine-tuning failed to raise the composite (19.7 ± 0.4 percent), while the neurosteroid analogue produced a similar figure (19.9 ± 0.5 percent). Thus only the ketamine condition yielded a net gain over baseline.

Table 1: Post-Treatment Accuracy (Mean ± SD Across Three Seeds)

Treatment	ARC-Easy (%)	LAMBADA (%)	Composite (%)
Untreated (pruned)	22.5 ± 0.0	20.0 ± 0.0	21.2 ± 0.0
Ketamine-like	32.3 ± 2.7	14.3 ± 1.2	23.3 ± 1.9
SSRI-like	22.3 ± 0.6	17.0 ± 0.8	19.7 ± 0.4
Neurosteroid-like	28.5 ± 1.1	11.3 ± 0.5	19.9 ± 0.5

Recovery, resistance to acute relapse, and cost-effectiveness

Relative to the untreated model, ketamine-like tuning provided a mean recovery of +2.1 percent (Table 2), whereas SSRI- and neurosteroid-like procedures posted small negative shifts (–1.6 percent and –1.3 percent respectively). When a further 30 percent of weights were removed to simulate sudden relapse, the ketamine variant lost just 1.5 percent of accuracy and retained a marginal net benefit. Efficiency, defined as recovery per petaFLOP, was positive only for the ketamine analogue (≈15), confirming that structural rebuilding delivered the best return on compute (Table 3).

Table 2: Treatment Effects (Mean ± SD Across Seeds)

Treatment	Recovery (%)	Relapse Drop (%)	Net Retained (%)
Ketamine-like	+2.1 ± 1.9	1.5 ± 0.7	+0.6
SSRI-like	-1.6 ± 0.4	-2.5 ± 0.7	+0.9
Neurosteroid-like	-1.3 ± 0.5	-0.9 ± 2.3	-0.4

Table 3: Isodose Efficiency Comparison

Treatment	Recovery (%)	FLOPs	Efficiency (Recovery/PetaFLOP)
Ketamine-like	+2.1	1.35×10^{14}	15.40 ± 13.69
SSRI-like	-1.6	1.35×10^{14}	-11.70 ± 3.14
Neurosteroid-like	-1.3	1.29×10^{14}	-10.35 ± 3.99

Durability under progressive stress

Eight rounds of incremental 10 percent pruning, each followed by short maintenance fine-tuning, eroded initial gains (Table 4). The ketamine model began cycle 0 at 27.3 percent on ARC-Easy but fell by 10 points over the series. The SSRI mechanism, despite its modest start (21.3 percent), declined by only 3.3 points, finishing near 18 percent—comparable to the other two conditions, which

converged in the same band by cycle 8. Hence ketamine produced the strongest early recovery but was most vulnerable to cumulative damage.

Table 4: Longitudinal ARC-Easy Accuracy (Mean ± SD Across Seeds)

Cycle	Ketamine-like (%)	SSRI-like (%)	Neurosteroid-like (%)
0	27.3 ± 3.8	21.3 ± 1.9	25.3 ± 1.9
1	19.3 ± 2.5	24.0 ± 0.0	30.7 ± 2.5
2	16.7 ± 2.5	5.3 ± 0.9	4.7 ± 2.5
3	19.3 ± 3.4	12.0 ± 0.0	29.3 ± 0.9
4	22.0 ± 3.3	8.0 ± 0.0	28.0 ± 1.6
5	7.3 ± 0.9	30.0 ± 0.0	7.3 ± 0.9
6	16.7 ± 1.9	4.0 ± 0.0	18.7 ± 0.9
7	20.0 ± 0.0	24.7 ± 3.4	22.0 ± 1.6
8	17.3 ± 1.9	18.0 ± 0.0	18.0 ± 0.0

Optimising ketamine-like parameters

Table 5: Isodose Sweep Rankings (Single Seed, by Composite Score)

Rank	LoRA r	Epochs	Recovery (%)	Efficiency	Relapse Drop (%)	Composite Score
1	32	6	+3.5	25.87	3.0	6.79
2	64	3	+0.0	0.00	0.5	3.80
3	96	2	-2.5	-18.48	0.0	1.15
4	128	2	-4.5	-24.95	-2.0	0.51
5	16	12	-2.0	-14.78	5.5	-0.48

An isodose sweep varying LoRA rank identified r = 32 (α scaled accordingly, six epochs) as the sweet spot (Table 5). In a single-seed screen this setting improved composite accuracy by 3.5 percent and

yielded the highest efficiency. Replication under three seeds confirmed a mean recovery of 1.5 ± 1.4 percent, still leading the other ranks.

Transfer to alternate pruned architectures

Applying the optimised r = 32 recipe to a 40 percent-pruned Gemma-2-2B increased its composite score by 5.0 ± 1.5 percent, and to a six-layer-pruned Mistral-7B by 3.5 ± 1.2 percent (Table 6). Implementation constraints prevented evaluation on the depth-pruned Phi-3 model. Averaged across the two successful transfers, recovery was 4.0 ± 0.7 percent, suggesting that the structurally focused protocol generalises beyond the original Llama substrate.

Table 6: Generalization Results (Optimal Rank 32, Mean ± SD Across Three Seeds Where Applicable)

Model	Baseline Composite (%)	Recovery (%)	Efficiency (Mean)	Relapse Drop (Mean %)	Delta vs. Primary (%)
Primary: Llama-3.2-1B (60% pruned)	22.5	+3.5	25.87	3.0	–
Gemma-2-2B (40% MLP pruned)	32.0	+5.0 ± 1.5	36.96	6.0	+1.5
Mistral-7B (6 layers removed)	67.5	+3.5 ± 1.2	25.87	1.3	+0.0

Discussion

Mechanism-specific recovery patterns

Placing the three adaptation strategies under an identical compute ceiling made their qualitative differences unmistakable. The structural-rebuilding, ketamine-like procedure restored the largest share of lost accuracy, confirming that a broad reinstatement of projection pathways can quickly patch the knowledge gaps created by aggressive width pruning [3]. The same intervention also showed the best

short-term resilience: when another 30 percent of weights were removed, performance slipped only slightly. These advantages were obtained with the lowest cost per percentage point recovered, underscoring that, in an acutely fragile network, re-expanding capacity in a targeted way is a highly productive use of FLOPs.

The monoaminergic, SSRI-like schedule told a different story. Immediate gains were negligible, yet the model deteriorated least during eight rounds of progressive pruning. A narrow, low-rank update to attention maps therefore seems insufficient for acute repair but helpful for long-horizon stability—mirroring the delayed clinical efficacy typically associated with selective-serotonin agents. The neurosteroid analogue split the difference: balanced early gains but a steeper late-cycle slide, consistent with a mechanism that tempers activation noise without rebuilding lost structure.

Taken together, these results indicate that pruning does not simply lower performance in a uniform way; rather, it exposes weaknesses that different adaptation styles exploit or leave untouched. Because compute budgets were equalised, the outcome spread must arise from mechanism, not dose.

Optimal scale for structural regrowth

Sweeping the rank parameter inside the ketamine-style LoRA layers revealed a clear peak at rank 32. Larger ranks added little and sometimes eroded accuracy, implying that beyond a moderate width the extra capacity re-introduces noisy weights instead of reviving useful representations. This observation dovetails with the "width-pruning dichotomy" reported by Martra [3], where over-expansion after pruning can degrade factual recall.

Crucially, the same rank-32 recipe travelled well. Applied unaltered to a 40 percent-pruned Gemma-2-2B and a depth-pruned Mistral-7B, it delivered comparable or better recovery, even though those backbones differ in size and pruning style. Generalisation of this kind suggests that moderate structural regrowth targets a common failure mode shared by sparsified transformers.

Implications for fragility and repair

The behavioural split uncovered here—rapid but erosion-prone rebuilding versus slower, steadier refinement—parallels therapeutic trade-offs seen in neuropsychiatry, where fast-acting glutamatergic agents and slower monoaminergic agents are combined for complementary benefit. In language models, a similar combination might pair a brief, high-rank regrowth phase to plug the worst factual holes with a low-rank maintenance phase to preserve gains during continued sparsification or domain drift.

More broadly, the data reinforce the notion that pruning enforces a selective filter rather than a universal handicap. When that filter removes too much of the knowledge substrate, carefully chosen low-rank updates can restore useful pathways without reconstructing the whole network—provided the update scale is matched to the remaining capacity.

Therapeutic analogies suggested by the simulation

Although every experiment took place inside a purely computational landscape, several themes echo long-standing ideas about synaptic loss in major depressive disorder [4]. Heavy width pruning created a "latent-deficit" network that functioned acceptably until challenged, much as chronically stressed corticolimbic circuits appear outwardly intact yet decompensate under additional load.

Within that fragile substrate, the ketamine-like adaptation delivered the fastest and most compute-efficient restoration, mirroring the rapid symptom relief seen with glutamatergic agents in treatment-resistant depression [12]. The discovery that a moderate LoRA rank (≈32) maximised benefit without overshoot brings to mind clinical work showing sub-anaesthetic ketamine doses can trigger synaptogenesis through BDNF-mTOR signalling while avoiding adverse dissociation [13]. Larger ranks failed to help and occasionally harmed performance, echoing reports that repeated high-dose infusions yield diminishing returns.

By contrast, the SSRI-like schedule offered negligible immediate help but protected the model during eight stress cycles, a pattern reminiscent of monoaminergic drugs that accumulate benefit slowly and reduce long-term relapse risk in milder illness courses. The neurosteroid analogue stabilised the

network quickly yet became vulnerable later, paralleling GABAergic modulators whose clinical efficacy can wane once dosing stops if structural deficits remain. Because each regimen consumed an equal compute "budget," these differences must arise from the way each algorithm interacts with a sparsified architecture, not from unequal dosing.

Broader implications for pruning fragility

The results reinforce the view that pruning acts as a selective filter rather than a uniform lesion [3]. Structural regrowth re-opens dormant pathways and is immediately rewarding, but unless followed by lower-intensity optimisation it remains susceptible to further damage. Functional fine-tuning alone struggles to repair deep gaps yet keeps whatever remains more robust over time. This complementarity resembles combination strategies now common in psychiatry (Table 7), where an acute plasticity enhancer (e.g., ketamine) is paired with a maintenance agent (e.g., SSRI) to sustain remission.

The fact that the moderate-rank ketamine recipe generalised to Gemma and Mistral backbones suggests that pruning-related vulnerability may obey common geometric constraints across transformer families. If so, modest structural "boosters" could become a generic remedy for heavily compressed models, just as ketamine is being explored across diagnostic categories marked by synaptic loss.

Novel contribution and practical relevance

By transplanting Cheung's pruning-plasticity thought experiment from abstract feed-forward nets into real, publicly available transformers, this study delivers the first head-to-head, compute-matched test of three biologically inspired recovery motifs. Using LoRA adapters on models that already exhibit the "width-pruning dichotomy"—knowledge loss paired with crisper instruction following [3]—we show that structural rebuilding, gradual refinement, and tonic inhibition do not merely differ in degree but in the kind of capability they restore. The isodose design is central: equal FLOP budgets rule out

the common objection that one method simply "trains longer," allowing genuine mechanistic contrasts to surface.

Table 7: LLM Parameter Terms Translated to Human Depression Treatment Analogies

LLM Term	Human Depression Analogy	Explanation & Notes
Pruning (e.g., 60% MLP removal)	Excessive synaptic/dendritic spine pruning from chronic stress	Stress-driven microglial over-pruning reduces connectivity, creating latent vulnerability unmasked by triggers (e.g., anhedonia/cognitive deficits in TRD). Note: Model pruning is magnitude-based and static; human pruning involves inflammation and is reversible but slow.
Fragility/Baseline Drop	Subclinical synaptic deficit → symptoms under stress	Daily functioning preserved until stressors reveal breakdown (e.g., relapse in remitted MDD). Note: Benchmarks like ARC-Easy/LAMBADA probe knowledge/reasoning; human equivalents include cognitive tests showing blunting.
LoRA Adaptation	Targeted neuroplasticity enhancement (e.g., synaptogenesis)	Adds "new pathways" efficiently → ketamine's rapid BDNF/TrkB-driven spine regrowth. Note: LoRA freezes base weights; human plasticity interacts with existing circuits dynamically.
LoRA Rank (optimal ~32)	Extent of new synaptic connections (dose-dependent plasticity)	Moderate rank = balanced regrowth → "Goldilocks" ketamine dose for sufficient spines without overload. High rank risks diminishing returns (like dissociation/tolerance); low rank = undertreatment. Note: Optimal moderate aligns with spaced, sub-anesthetic protocols avoiding side effects.
Alpha (scaling)	Amplification of plasticity signals (e.g., mTOR/BDNF strength)	Boosts regrowth impact → stronger ketamine "burst" but potential over-excitability. Note: Mirrors pathway hyperactivity risks in high-dose trials.
Epochs (iterations)	Duration/exposures of treatment sessions	More epochs = repeated infusions → spaced dosing in ketamine's short plasticity window (2-3 days) to sustain effects. Note: Optimal ~6 suggests limited sessions prevent tolerance.
FLOPs/Isodose	Therapeutic dose (concentration × exposure)	Equated cost = bioequivalent comparisons (e.g., ketamine vs. esketamine). Note: Ensures fair mechanism testing, isolating "drug" effects from dosing.
Recovery/Composite Score	Antidepressant response/remission	Post-treatment gain → rapid symptom relief (ketamine: high response in TRD). Note: Modest model gains (2-5%) vs. clinical 50-70% highlight scale differences.
Efficiency (per FLOPs)	Response per dose (minimizing side effects)	High at optimal rank → best outcomes with least exposure/risk. Note: Supports personalized dosing to avoid non-response.
Relapse Drop (post-stress)	Relapse vulnerability under new stressors	Low drop → durable remission via synaptic "reserve." Note: Ketamine boosters in clinics extend effects.
Generalization Across Models	Efficacy across patient subtypes/comorbidities	Optimal dose transfers → ketamine's broad use (MDD, PTSD). Note: Low variability suggests biomarker-guided selection (e.g., inflammation markers).

Notes on Limitations and Speculative Nature: *These parallels are heuristic only; LLMs lack emotions, homeostasis, or glia-neuron interactions central to depression. Model "recovery" is task-specific and modest; human outcomes vary widely due to genetics, environment, and subjectivity. Farfetched aspects: Direct equivalence of LoRA rank to synaptic count ignores qualitative differences (e.g., excitatory vs. inhibitory balance). Use for hypothesis generation, not clinical inference.*

A second advance is the systematic sweep of the structural (ketamine-like) setting. The data isolate a moderate adapter rank that maximises accuracy per petaFLOP and, importantly, travels unchanged to other sparsified backbones. In practice, compression pipelines that sacrifice parameters for latency or memory [1,2] could insert such a targeted regrowth pass to regain robustness without surrendering efficiency. Conceptually, the finding mirrors clinical discussions that ketamine's benefit comes from a brief, balanced plasticity window rather than from maximal or chronic exposure [12].

Caveats and boundary conditions

Several limitations counsel restraint. All trials used a single one-billion-parameter Llama derivative pruned extremely hard; larger or more lightly compressed models may respond differently. The tuning set focused on instruction compliance, so factual repair may be under-represented. Long-term stress was modelled with crude iterative pruning and short "maintenance" bursts; real deployment drifts—and biological stressors—are far more nuanced.

Mechanistic symmetry was also imperfect by design: the SSRI-like adapter touched only query and value projections, whereas the ketamine analogue modified every projection layer. These choices embody biological intuition, yet they also mean that pure ablation of "rank" versus "layer coverage" effects is incomplete. Transfer tests succeeded on Gemma and Mistral but not on the depth-pruned Phi-3, leaving that axis partly unexplored. Finally, any neuroscientific analogy must be read as heuristic; transformers lack recurrence, cell-type diversity, and affect—crucial ingredients of mood circuits.

Outlook

Within those bounds, moderate-sized structural regrowth stands out as the fastest and most compute-efficient route to stabilising heavily pruned transformers, while slower low-rank refinement offers durability once crisis passes. Scaling the protocol to multi-billion-parameter models, weaving in alignment and safety objectives, and experimenting with staged or hybrid schedules are promising next steps toward compression strategies that keep both resource use and capability in balance.

References

[1] Frantar E, et al. Sparsegpt: Massive language models can be accurately pruned in one-shot. Proc Int Conf Mach Learn. 2023; p. 10323–10337.

[2] Sun M, et al. A simple and effective pruning approach for large language models. arXiv:2306.11695. 2023.

[3] Martra P. Fragile Knowledge, Robust Instruction-Following: The Width Pruning Dichotomy in Llama-3.2. arXiv:2512.22671. 2025.

[4] Duman RS, et al. Synaptic dysfunction in depression: Potential therapeutic targets. Science. 2012;338(6103):68–72.

[5] Cheung N. Structural rebuilding confers superior long-term resilience: A unified multi-mechanism computational comparison of antidepressants in chronic stress. Zenodo. 2026. https://doi.org/10.5281/zenodo.18295517

[6] Hu EJ, et al. Lora: Low-rank adaptation of large language models. ICLR. 2022;1(2):3.

[7] Meta. Llama 3.2 release announcement. AI at Meta. 2024. https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices

[8] Wolf T, et al. Transformers: State-of-the-art natural language processing. Proc Conf Empir Methods Nat Lang Process. 2020; p. 38–45.

[9] Databricks. databricks-dolly-15k [Data set]. Hugging Face. 2023. https://huggingface.co/datasets/databricks/databricks-dolly-15k

[10] Clark P, et al. Think you have solved question answering? try arc, the ai2 reasoning challenge. arXiv:1803.05457. 2018.

[11] Paperno D, et al. The LAMBADA dataset: Word prediction requiring a broad discourse context. Proc 54th Annu Meet Assoc Comput Linguist. 2016; p. 1525–1534.

[12] Krystal JH, et al. Ketamine: A paradigm shift for depression research and treatment. Neuron. 2019;101:774–778.

[13] Abdallah CG, et al. Ketamine's mechanism of action: a path to rapid‐acting antidepressants. Depression and anxiety. 2016;33(8):689–697.