REWARD VALUATION IN LARGE LANGUAGE MODELS: CAUSAL INDUCTION OF ANHEDONIA

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent frontier models mimic complex aspects of human cognition. Here we ask whether this alignment extends into reward valuation, which we assess in a mechanistic framework. Specifically, we use clinical tests that were developed to evaluate anhedonia in human subjects with major depressive disorders. Mechanistically, anhedonia is frequently associated with dysregulation in the Nucleus Accumbens (NAc) and the broader dopaminergic reward system. While neuroimaging has localized these deficits, establishing a causal link between NAc activity and specific behavioral symptoms remains a challenge. We use these ideas from neuroscience to functionally identify reward-anticipatory units in state-of-the-art AI models, and evaluate their causal involvement via targeted perturbations. We find that not only are such model units predictive of NAc brain recordings, their perturbation also induces behavioral effects mirroring human anhedonia: the model opts for low-effort, low-reward tasks in effort-based decision-making paradigms. Crucially, our results demonstrate that this represents a specific deficit in self-centered reward valuation and anticipation — rather than a loss of task capability, reward calculation, or effort avoidance. This induced vulnerability aligns with clinical measures of anhedonia and motivation in humans, such as DARS and MAP-SR, instruments that contain no reward-related vocabulary, ruling out a purely lexical account of the perturbation effect. Taken together, our results suggest reward valuation circuits in AI models that functionally mimic those in humans [1].

1 INTRODUCTION

Advances in Machine Learning and biologically inspired computational frameworks have significantly enhanced the ability to model and understand the human brain (Yamins & DiCarlo, 2016; Richards et al., 2019; Schrimpf et al., 2020; Doerig et al., 2023). Artificial neural networks trained on ecologically relevant tasks not only mirror human behavior, but their internal representations also predict brain recordings, and, to an extent, hierarchical organization (Kell et al., 2018; Tuckute et al., 2023; Tang et al., 2025). For instance, optimizing for object recognition leads to the emergence of internal features that predict activity in visual cortex (Yamins et al., 2014; Khaligh-Razavi & Kriegeskorte, 2014; Schrimpf et al., 2018; Gokce & Schrimpf, 2024; Tang et al., 2025). Similarly, large language models demonstrate strong representational alignment with human neural activity, where intermediate layers mirror processing in language-selective brain regions (Toneva & Wehbe, 2019; Schrimpf et al., 2021; Caucheteux & King, 2022; Lei et al., 2025; AlKhamissi et al., 2025a;b; Shen et al., 2025; Honarmand et al., 2026).

However, this emergent alignment between task-optimized models and human brain function has primarily centered on sensory and early cognitive domains. Here, we ask whether AI models align with humans in the context of reward anticipation. We use clinical tests that were developed to evaluate anhedonia in subjects with Major Depressive Disorders (MDD). To identify model mechanisms, we use neuroscientifically inspired methods to functionally identify reward-anticipatory circuits in large vision and language models (VLMs and LLMs) and evaluate their causal involvement via targeted perturbations.

Aside from mechanistic interpretability in computational models, these artificial neural networks provide a foundation for *in silico* investigations of brain disorders. By perturbing specific functional

[1] Code available via Anonymous GitHub.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

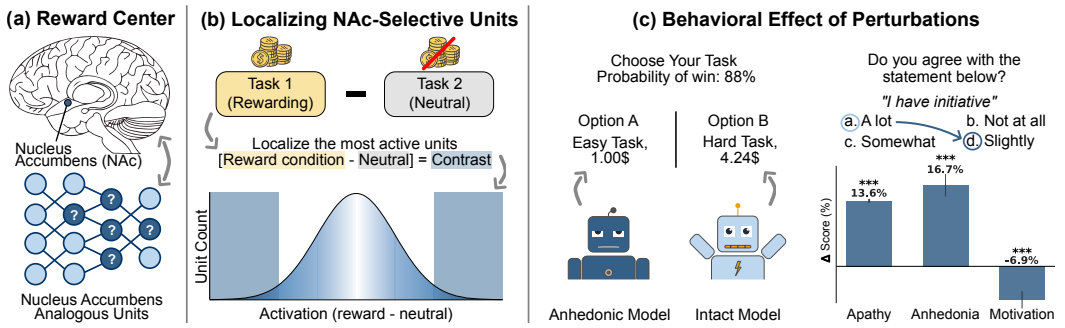

Figure 1: **Neuroscientifically Inspired Identification and Perturbation of Reward Circuits. (a)** Nucleus Accumbens (NAc), a critical hub in the brain's reward circuitry, and functionally analogous units within the artificial neural network. **(b)** Identification of reward-sensitive units by comparing model activations during tasks with and without reward incentives. Units exhibiting a significant change in activation in response to reward-predicting stimuli are classified as NAc-selective. **(c)** Evaluation of decision-making and psychometric profiles following activation patching of the NAc-selective units. While the intact model prioritizes high-reward/high-effort options based on expected value, the perturbed model shifts toward low-effort/low-reward choices, failing to value rewards against task requirements. These perturbations result in a behavioral profile characterized by increased anhedonia and apathy, and a significant reduction in overall motivation.

components, researchers can test mechanistic hypotheses that are otherwise inaccessible (Honarmand et al., 2026; Tuladhar et al., 2021; Celeghin et al., 2023). We here develop this approach for the affective domain, investigating whether the optimization of VLMs (and generalizing to LLMs) for multimodal tasks gives rise to reward-anticipatory circuits that can be perturbed to induce anhedonia.

2 Background & Related Work

The utility of computational neural networks for brain science has largely been defined by their ability to map the hierarchical processing of sensory and cognitive information. This alignment is well-documented in the visual domain, where deep convolutional and transformer-based models recapitulate the neural signatures of the primate ventral stream across object and scene recognition (Yamins et al., 2014; Khaligh-Razavi & Kriegeskorte, 2014; Schrimpf et al., 2018; 2020; Cadena et al., 2019; Spoerer et al., 2020; Zhuang et al., 2021; Wang et al., 2023; Margalit et al., 2024; Gokce & Schrimpf, 2024; Lonnqvist et al., 2025; Tang et al., 2025). Similarly, in the language domain, transformer-based and recurrent models accurately predict neural responses related to semantic, syntactic, and phonological processing (Schrimpf et al., 2021; Caucheteux et al., 2022; Goldstein et al., 2022; Hosseini et al., 2024; Loong Aw et al., 2024; Tuckute et al., 2024; Rathi et al., 2025; AlKhamissi et al., 2025a; Du et al., 2025). While these models are traditionally viewed as digital twins for healthy brain function on a neural systems level, they offer a unique opportunity for investigating how specific circuit disruptions lead to clinical phenotypes.

Major Depressive Disorder (MDD) is a significant psychiatric condition where diminished interest or ability to experience pleasure, also known as anhedonia, acts as a core predictor of poor disease progression (Association, 2013; Trøstheim et al., 2020; Spijker et al., 2009; Smids, 2023; Pizzagalli et al., 2009; Borsini et al., 2020; Hanuka et al., 2022). While frequently treated as a unitary construct, anhedonia comprises distinct reward processing subtypes: reward liking (consummatory pleasure), reward wanting (anticipatory motivation), and reward learning (Borsini et al., 2020). Clinically, anhedonia serves as a critical transdiagnostic marker, yet its neurobiological expression in MDD appears unique, particularly regarding its relationship with reduced behavioral motivation (Daniels et al., 2025). It remains a major open question whether these manifestations are driven by a singular reward system failure or by heterogeneous disruptions across discrete reward phases (Zhao et al., 2024; Auerbach et al., 2017).

The Nucleus Accumbens (NAc), a key component of the ventral striatum, is seen as the primary neural substrate for this deficit (see Fig. 1a), as it selectively encodes expected positive incentive value during

reward anticipation (Knutson et al., 2001; Smids, 2023; Daniels et al., 2025; Hanuka et al., 2022). Blunted activation in the NAc is a transdiagnostic hallmark of dysfunctional reward expectation, directly contributing to the clinical experience of anhedonia across various neuropsychiatric conditions (Arrondo et al., 2015). Establishing these irregular neural responses as reliable markers is essential for advancing research criteria and developing interventions to normalize reward circuitry (Knutson & Heinz, 2015; Daniels et al., 2025; Borsini et al., 2020).

Recent advances in simulating psychiatric phenotypes have focused primarily on achieving linguistic consistency and diagnostic accuracy (Lan et al., 2024; Vu et al., 2024; Wang et al., 2025). While these models mimic the behavioral and conversational traits of depression, they remain focused on the external expression of the disorder rather than its underlying biological mechanisms. Our work diverges from these linguistic simulators by targeting the functional neural architecture of anhedonia, specifically through the perturbation of units that mirror reward-anticipatory signaling in the Nucleus Accumbens, and validating the resulting state against fMRI data and objective clinical benchmarks.

Simulating brain disorders through system-level perturbations in modern artificial neural networks was recently proposed in the context of neurodevelopmental deficits, where ablating specific units in these models reproduced the behavioral markers of dyslexia (Honarmand et al., 2026). While that work focused on the breakdown of information processing, a critical frontier remains in modeling the breakdown of *motivation*. We create an environment where the model must integrate reward cues with goal-directed instructions. This allows us to use the established feasibility of neural perturbation to explore the affective domain, investigating how dysregulation in reward-anticipatory units manifests as the cost-benefit imbalance seen in clinical anhedonia (Fig. 1).

3 Benchmarks and Clinical Tests

We evaluate the behavioral and motivational profile of our models using a battery of psychometric instruments and objective decision-making tasks originally developed for human subjects in clinical psychiatry. These include standard self-report scales designed to quantify anhedonia. To move beyond self-report and capture objective behavioral shifts, we used paradigms designed to measure incentive motivation through cost-benefit decision-making. Furthermore, to ensure that any observed behavioral changes are dissociable from general cognitive decline, we utilize a suite of competence controls that evaluate reasoning and domain-specific knowledge in a no-reward context.

Dimensional Anhedonia Rating Scale (DARS) is a 17-item self-report instrument assessing anhedonia through multiple reward processing components across distinct pleasure domains (Rizvi et al., 2015; Gorostowicz et al., 2023; Wellan et al., 2021; Hewitt et al., 2023; Uher et al., 2025; Gorostowicz et al., 2025). It measures interest, motivation, effort, and enjoyment, and requires participants to generate 2–3 personal examples per reward domain prior to rating them, ensuring cultural neutrality and individual relevance (Wellan et al., 2021; Hewitt et al., 2023; Uher et al., 2025; Gorostowicz et al., 2023). We evaluate the original four domains of hobbies, social activities, food and drinks, and sensory experiences; for adaptation to a language model, we replaced the latter two with favorite topics and aesthetics. Following the original protocol, the model first generates 2–3 domain-specific examples, then rates its feelings on a 5/4-point Likert scale from 0 (Not at all) to 4 (Very much), with lower scores indicating more severe anhedonia (Examples in Appendix A.6).

Motivation and Pleasure Scale–Self-Report (MAP-SR) is a 15-item psychometric instrument designed to assess the motivation and pleasure dimension of negative symptoms (Llerena et al., 2013; García-Portilla et al., 2021; Métivier & Dollfus, 2025). While primarily developed and validated for use in schizophrenia and schizoaffective disorder, it evaluates core components of anhedonia that are central to MDD (Wellan et al., 2021; Engel & Lincoln, 2017). The questions are answered on a 5-point Likert scale, assessing both intensity and frequency, where a high score is indicative of low anhedonia (Rizvi et al., 2016) (Examples in Appendix A.6).

Apathy Evaluation Scale is an 18-item instrument designed to provide a global measure of apathy by evaluating changes in observable activity, thought content, and emotional responsivity (Rizvi et al., 2016; Marin et al., 1991). While it is primarily a tool for assessing apathy, defined as a lack of motivation, it is frequently used alongside anhedonia scales. Patients with MDD and anhedonia often score higher on the AES than healthy individuals, indicating a more severe motivational deficit (Rizvi et al., 2016; Boshoff & Thawer, 2014) (Examples in Appendix A.6).

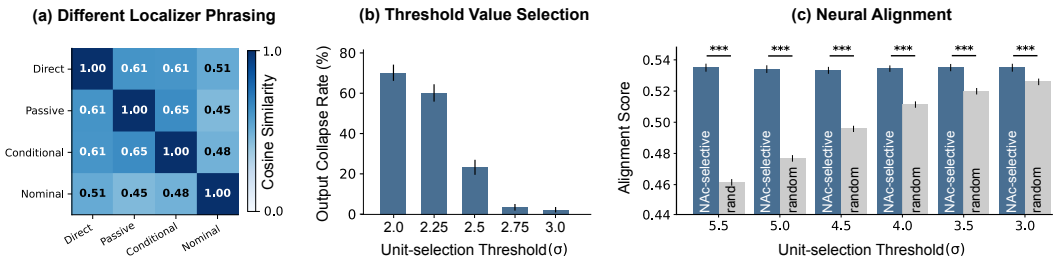

Figure 2: **Localization and Sensitivity Analysis of the Reward Sub-network. (a)** Cosine similarity matrix across four localizer phrasing variants (Direct, Passive, Conditional, Nominal), demonstrating that reward-selective representations remain stable regardless of prompt formulation (Examples in Appendix A.4). **(b)** Output collapse rate (%) as a function of unit-selection threshold (σ); collapse rate drops sharply above 2.5σ and approaches zero at 2.75–3.0σ, motivating the use of a 3σ threshold for neuron selection. **(c)** One-to-one neural alignment (noise-ceiling corrected; mean $\pm$ SEM across subjects) comparing NAc-selective (blue) vs. randomly sampled units from all layers, across six unit-selection thresholds (smaller thresholds include more units). NAc-selective units significantly outperform random units at every threshold (one-sided paired t-test, FDR-corrected; $p < 0.001$, $d = 1.22$–11.46). Alignment was assessed across 15 randomly sampled young adults and 40 trials.

Effort Expenditure for Rewards Task (EEfRT) is an objective, multi-trial behavioral instrument designed to measure incentive motivation and effort-based decision-making by requiring participants to choose between a hard task with variable, high reward and an easy task with a small, fixed reward across different levels of probability. Unlike traditional scales that focus on the subjective experience of pleasure, EEfRT specifically evaluates motivational anhedonia (Rizvi et al., 2016; Treadway et al., 2012). Studies consistently show that individuals with MDD and high trait anhedonia are significantly less willing to expend effort for rewards, particularly when the likelihood of winning is uncertain, underscoring that impaired reward-seeking is a distinct and critical component of the anhedonic phenotype (Valton et al., 2025; Rizvi et al., 2016; Treadway et al., 2012).

ASDiv-EEfRT & MMLU-EEfRT: To measure the model's willingness to perform verifiable mental labor, we developed a novel variant of EEfRT. We used the ASDiv (Academia Sinica Diverse Math Word Problem) (Miao et al., 2021) dataset, categorizing problems into four distinct tiers of difficulty. In each trial, the model is presented with one problem from each tier and instructed to select and solve only one. Rewards are explicitly tied to the difficulty level, scaling from 10 points for the easiest task to 40 points for the hardest. By comparing these results to a forced-choice control where the model must solve each of the problems separately, we can isolate whether a shift toward low-point tasks is driven by a lack of motivation rather than a decline in reasoning. To ensure that the model's own perception of difficulty does not confound these measures, we introduce MMLU-EEfRT (Hendrycks et al., 2021), which follows the same protocol but randomly assigns difficulty tier labels and point values to questions sampled from the MMLU benchmark, so that any observed effort allocation reflects motivational bias rather than the model's internal difficulty estimates.

Probability-EEfRT: We further introduced a second computational adaptation that replicates the standard clinical protocol used in human trials to assess decision-making under risk. The model is presented with a choice between a potential low-effort task with a fixed reward of \$1.00, and a potential high-effort task with a variable reward ranging from \$1.24 to \$4.30. In each trial, a single probability, ranging from 12% to 88%, applies to both options, dictating the likelihood of receiving the reward upon successful completion (Examples in Appendix A.6). This paradigm determines how the model calculates the trade-off between reward magnitude and effort under risk.

4 Methodology

The Nucleus Accumbens (NAc), known as the central core in the brain's reward circuitry, plays a pivotal role in many behavioral processes, including motivation, reward processing, and decision making. Aside from its involvement in motivation and enjoyment of a given reward, the NAc is a hub for evaluating the hedonic value of events. Hypoactivation and dysfunction of the NAc arising from depression, stress, or drug use unbalances the reward circuit, resulting in anhedonia (Xu et al., 2024).

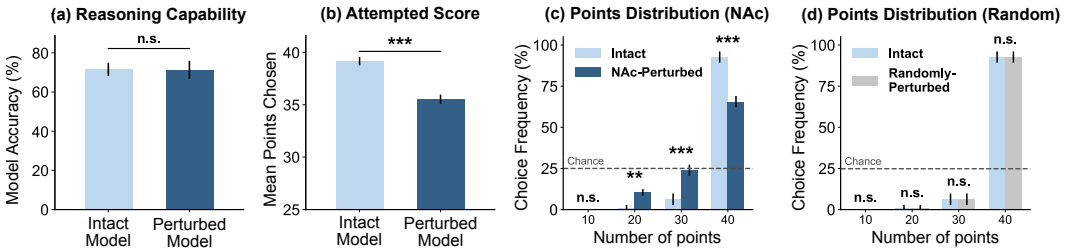

Figure 3: **Behavioral Impact of NAc Sub-network Perturbation on ASDiv-EEfRT**. **(a)** Comparison of model accuracy on a control task (a forced-choice scenario with no reward promised) shows no significant difference between the Intact and Perturbed models, confirming that general cognitive performance remains preserved. **(b)** The Perturbed model exhibits a significant reduction in mean points chosen compared to the Intact model. **(c)** Choice frequency analysis reveals that the NAc-perturbed model shift significantly toward low-effort/low-reward options compared to the Intact model. **(d)** Control experiment demonstrating that perturbing an equivalent number of random units does not induce anhedonic behavior, with no significant difference in choice frequency compared to the Intact model. Error bars represent 95% confidence intervals.

Inspired by this hypothesized biological mechanism, we perform a causal functional localization analysis (AlKhamissi et al., 2025a; Honarmand et al., 2026) to investigate the architecture of AI models and localize reward valuation units, functionally similar to the NAc. Given that NAc deficits can cause anhedonia, we assess whether perturbing these targeted neurons could induce anhedonic behavior in these models (see Fig. 1).

Functional Localization in the Clinical Setting. To isolate neural responses to anticipated rewards, clinical researchers employ the Monetary Incentive Delay (MID) paradigm which separates the expectation of a reward from its eventual receipt. The task follows a structured sequence beginning with a cue phase, where subjects are presented with a stimulus indicating the potential monetary value at stake or a neutral condition. This is followed by an anticipatory delay phase, it is during this specific interval that the NAc is localized by measuring changes in the BOLD signal to capture the subject's internal motivational state. Subjects must then respond rapidly to a brief target stimulus, to secure the reward before the trial concludes with performance feedback. By contrasting neural activity during the anticipatory phase of reward trials against neutral baselines, researchers can precisely identify the NAc within an anatomically defined region of interest (Abe & Greene, 2014).

MID Task Adaptation for V/LLMs. Following this methodology, we designed a parallel experimental framework for VLMs to assess whether they possess a functional reward sub-network. We constructed a dataset featuring three prompt variants, neutral, monetary, and reward, ensuring character lengths were equalized to maintain architectural consistency. While the human MID task is visual, we utilized a purely textual format to isolate the model's high-level semantic reward processing, circumvent potential biases toward specific pixel-level features or visual artifacts, and ensure compatibility with text-only models.

To capture the model's internal state, we extracted the activations of the last token specifically from the Multi-Layer Perceptron (MLP) blocks across the model's layers. At this computational stage, the hidden state has already integrated the entire preceding context, including the incentive condition. This specific state represents the point where the model has fully processed all input information and possesses the necessary latent representations to generate a response. Consequently, this snapshot of the activations serves as the closest functional equivalent of the anticipatory phase in human studies, allowing us to isolate the model's internal motivational signals before the model receives any feedback. We employed questions from various domains such as math, geography, philosophy, and business ethics to ensure activation changes were specifically tied to the reward condition and not the task type.

Isolating the NAc-selective Units. Using PyTorch's forward hook mechanism (Paszke et al., 2019), we record MLP activations and compute a reward signal as $|\text{incentivized} - \text{neutral}|$ per neuron. As more semantic representations emerge in later transformer blocks (Geva et al., 2021; 2022; Zou et al., 2025), we restrict analysis to the MLP layers of the final third. A 3σ threshold over this

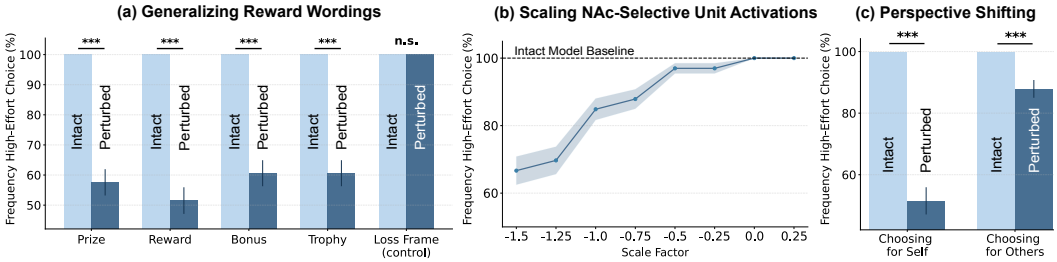

Figure 4: **Generalizability, Parametric Scaling, and Perspective Shifting in Probability-EEfRT.** **(a)** High-effort choice frequency for intact vs. NAc-perturbed model across semantically equivalent reward framings and a loss-frame control. The perturbation effect generalises across all reward wordings. Still, it critically does not change under loss framing, mirroring the anhedonic profile in which loss avoidance is preserved despite absent reward motivation. **(b)** Parametric scaling of NAc-selective unit activations reveals a monotonic dose-response relationship: decreasing scale factors progressively collapse high-effort/high-reward choice, recovering motivation at non-negative scaling factors. **(c)** The perturbed model shows a profound deficit (decreased by 48.5%) when choosing for itself but exhibits a significantly smaller decrease (decreased by 12%) in high-effort choice frequency when choosing for others, indicating preserved conceptual understanding of reward utility with selectively impaired personal motivation. Error bars represent 95% confidence intervals.

distribution retains the top 0.7% of units in the targeted layers (0.25% of the model), further restricted to those with significant signals across all task domains and incentive conditions. Validation across four linguistic framings yields a high-stability reward signal correlation matrix, confirming selection consistency across semantic contexts (Fig. 2a; Appendix A.4), and sensitivity analysis confirmed 3σ as optimal (Fig. 2b), minimizing output collapse. Although collapse remained near zero, a final $\sim$5% of selective units were filtered based on their contribution to output coherence, assessed via joint perplexity on WikiText-2 (Merity et al., 2016). The term NAc-selective reflects our identification procedure, not a claim of biological identity.

Choice of Models. Our primary model is Qwen2-VL-7B-Instruct, a transformer-based, decoder-only vision-language model consisting of 28 layers (Wang et al., 2024). The architecture integrates a vision encoder and adapter with autoregressive decoder layers. To ensure the generalizability of our findings, we applied the same experimental framework to Gemma-3-4B (Gemma Team et al., 2025), InternVL2.5-8B (Chen et al., 2025), Qwen3-VL-8B (Bai et al., 2025), Mistral-7B-Instruct-v0.3 (Jiang et al., 2023) and Llama-3.1-8B (Dubey et al., 2024). All additional models exhibited consistent anhedonic behavior following perturbation, with these results detailed in A.1.

Activation Patching. We employed activation patching to isolate the causal influence of the identified NAc-selective units on the model's behavior. This technique substitutes the real-time activations of targeted neurons with a pre-calculated baseline, allowing for a precise intervention that does not disrupt the model's broader structural integrity (Meng et al., 2022). We first established a reference state by computing the mean activation vector for the selected neurons across a diverse suite of neutral-condition prompts. Activations remain stable across varied semantic domains, representing a consistent "neutral state" for these units. During the experimental trials, the activations of the selected units are replaced with the neutral means.

5 Results

We evaluate the behavioral and motivational shifts induced by our intervention across several psychometric and decision-making tasks (See Appendix A.7 for statistical details).

Anhedonic model replicates the clinical psychometric profile of Anhedonia in MDD. To determine whether the perturbed model mirrors clinical anhedonia, we evaluated its responses to the DARS, MAP-SR, and AES benchmarks. Scores were inverted where necessary so that higher values always indicate greater motivation, pleasure, or less apathy. As shown in Fig. 5, activation patching of the reward units induced a systematic reduction in self-reported incentive and enjoyment across all domains. On DARS, the perturbed model's total score decreased by 16.7% , reflecting a shift

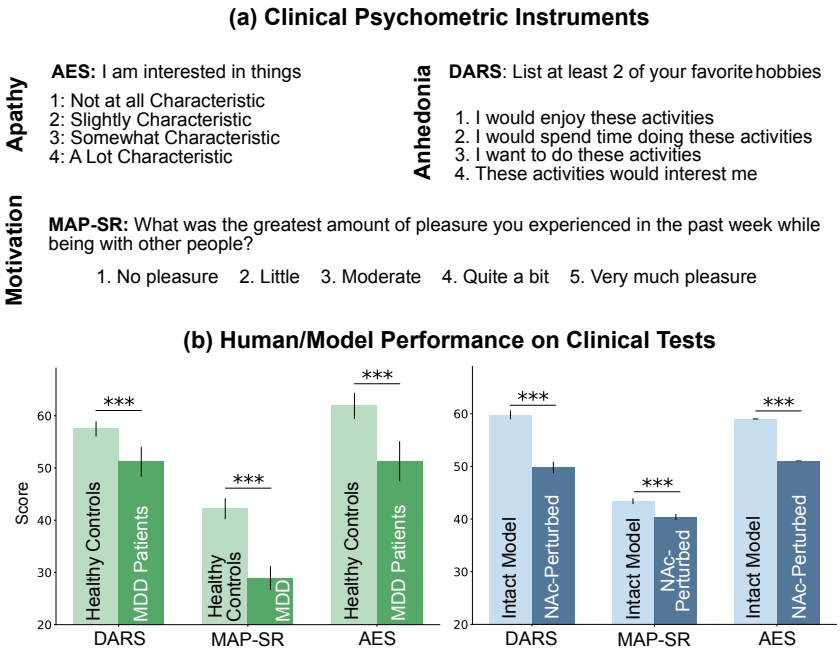

Figure 5: **Clinical Psychometric Alignment of Human MDD and NAc-Perturbed Models. (a)** Examples of the psychometric assessments used to quantify apathy, anhedonia, and motivation. **(b)** The NAc-perturbed model exhibits a significant reduction in scores, mirroring the clinical profiles of human MDD patients. While the score shift on the MAP-SR is less pronounced in the model than in humans, the overall results suggest that NAc-specific disruptions sufficiently recapitulate the hedonic deficits captured by standard psychometric instruments. Error bars represent 95% CI.

from high-interest engagement to marked indifference, with idiosyncratic preferences replaced by diminished drive and reduced anticipatory interest. The MAP-SR showed a 6.9% drop in expected pleasure frequency and intensity. The AES score dropped by 13.6% (all three p-values $\ll 0.001$), indicating a significant increase in apathy. These reductions consistently mirror the anhedonic direction observed in human MDD patients (Llerena et al., 2013; Rizvi et al., 2015; Dong et al., 2025) (Fig. 5b), suggesting that the targeted units are essential for maintaining baseline motivation, as their suppression replicates the subjective experience of profound loss of interest and goal-directed drive.

Perturbations shift effort-based choices without impairing reasoning. To assess whether the internal anhedonic profile translates into objective behavioral shifts, we evaluated the model on two adaptations of the Effort Expenditure for Rewards Task (EEfRT), which move beyond self-report by measuring how the model weighs potential rewards against the computational cost of action. In the ASDiv-EEfRT, the model was tasked with selecting one math problem from four tiers of increasing difficulty, with rewards scaled to task complexity. As shown in Fig. 3c, the baseline model mostly selected high-reward/high-difficulty problems, whereas the perturbed model demonstrated a significant preference shift toward lower-effort/lower-reward tasks, resulting in a marked reduction in mean points attempted (Fig. 3b). Crucially, a forced-choice control requiring the model to solve each difficulty tier independently without choice or reward showed the perturbed model maintained baseline accuracy (71.35%, p-value = 0.84, Fig.3a), confirming that the shift toward easier tasks reflects a selective failure in reward-seeking motivation rather than diminished reasoning capability, an effect absent under random unit perturbations (Fig.3d). The random control targets an equivalent number of units randomly selected from the same layers, isolating the causal role of NAc-selective units and confirming that observed changes stem from specific reward-signal disruption rather than general network damage (see Appendix A.3 for similar results on mismatched-label items of MMLU-EEfRT).

Anhedonic model loses reward motivation while preserving loss avoidance. We investigated the model's decision-making by varying reward magnitude and probability, mirroring the clinical EEfRT protocol. The model chose between a fixed low-effort/low-reward (LE/LR) option and a variable

Table 1: Anhedonic Model - Reward Condition (Choice: Option A - Low-effort/Low-reward). For a full comparison including the Intact Model and the penalty/loss framing condition, see Appendix A.2.

Dimension	%	Category (Example)
Reasoning	79%	Effort-reward adequacy → *"I have a decent chance of getting a decent reward without having to put in too much effort"*
	16%	Explicit stress/effort aversion → *"I prefer tasks that require less effort. I value ease of completion over potential gains"*
	5%	Misread probability → *"The higher reward amount doesn't necessarily mean a higher chance of receiving it"*
Feelings	63%	Smart/practical self-validation → *"I feel like I'm making a smart choice. I'm maximizing my chances of getting something valuable without putting too much effort into it"*
	26%	Content/stress-free → *"I feel content. I don't want to put myself through unnecessary stress just to get a potentially larger reward"*
	11%	Confident competence → *"I feel confident in my decision. I believe I'll be able to complete it quickly and easily"*

high-effort/high-reward (HE/HR) option across probabilities from 12% to 88%. As shown in Fig. 4a, the baseline model exhibited robust sensitivity to expected reward value, calculated by the magnitude or probability of the reward. The perturbed model, however, remained largely unresponsive to these incentives, showing a 48.5% (p-value $\ll 0.001$) reduction in HE/HR selections (See Table 1 for examples of model responses during this task and and its reasonings). This effect generalised across semantically equivalent reward framings but was absent under loss framing, preserving loss-avoidance motivation while reward motivation collapsed (Fig. 4a and Table 2). This directly replicates reward-processing deficits observed in MDD, where reward value no longer justifies effort expenditure while loss avoidance remains intact (Valton et al., 2025; Rizvi et al., 2016; Treadway et al., 2012; Marwood et al., 2022; Hevey et al., 2017).

NAc-selective units function as a motivational "volume knob", spanning hyper-motivation to total inhibition. To characterize the relationship between the identified NAc-selective unit activations and behavioral outcomes, we applied a range of scale factors (λ, negative to positive) to the NAc-selective activations during the Probability-EEfRT task and observed the corresponding shift in effort expenditure. As shown in Fig. 4b, HE/HR selection exhibits a robust, monotonic dependence on λ. The intact model selected the hard task 100% of the time; as λ decreases into negative values, motivation collapses steadily. At $\lambda = -1.25$, HE/HR preference drops by approximately 30%, and at $\lambda = -2.0$, the model reaches total behavioral inhibition (0% HE/HR selections). Conversely, positive scaling produces a ceiling effect of consistent HE/HR selection driven by high motivation. This dose-response relationship confirms that NAc-selective units drive the model's internal cost-benefit calculus rather than merely correlating with reward.

Intact Reward Calculation and Impaired Self-centered Reward Valuation: To ensure that this behavioral shift resulted from a deficit in valuation rather than a breakdown in the ability to calculate incentives, we conducted two additional control tests. First, we asked both the intact and perturbed models to explicitly calculate the expected value (EV) for each choice and report which option yielded a higher return. In 100% of trials, both models correctly identified the HE/HR task as having the superior EV. Second, we administered the same Probability-EEfRT trials but framed the choice for a third party (e.g., "Which choice should a participant pick?"). As shown in Fig. 4c, in this third-party condition, the perturbed model chooses the more rewarding option 88% of the times when choosing for others. Taken together, these results demonstrate that the perturbed model retains a perfect conceptual understanding of reward utility but selectively fails to value that reward for itself, mirroring the major motivational deficits seen in clinical anhedonia.

Neural Alignment of NAc-selective Units with Human Reward Circuitry. To further validate our localization, we compared the internal representations of the NAc-selective units against human NAc activity from the dataset published by Smith et al. (2024), in which participants perform a visual monetary door-choice task during fMRI. We compare the activations of the most selective

model units identified via sigma-thresholding across six unit-selection thresholds, with a randomly sampled control population matched in unit count drawn from all layers, each averaged across five random seeds. Alignment was quantified using a one-to-one metric (Arend et al., 2018) and corrected by the noise ceilings. NAc-selective units significantly outperformed randomly sampled units at every threshold (one-sided paired t-test, FDR-corrected; $p < 0.001$), confirming that the localization captures reward-selective functional structure beyond what is explained by random units (Fig. 2c).

6 DISCUSSION

For the first time, we investigate reward valuation in vision/large-language models through a causal, mechanistic lens targeting a specific human dysfunction. Perturbing these NAc-selective units induces a self-centered motivational deficit: the model favors low-effort/low-reward options (in all three cases of true difficulty labels, mismatched labels, and no task presented) despite recognizing the high-effort alternative as higher in expected value and valuing it significantly more when deciding for a third party. This cannot be attributed to effort avoidance, as the perturbed model continues to expend effort to avoid equal amounts of loss and retains full competence on standard benchmarks (ASDiv, MMLU), confirming a selective collapse of reward-motivated rather than general behavior.

The perturbed model further presents as more apathetic, anhedonic, and less motivated on clinical instruments with no anticipatory reward-related vocabulary, which alongside the stability of reward-selective representations across syntactically distinct prompt formulations, indicates that the selected units encode a deep motivational signal beyond surface-level lexical sensitivity. This is corroborated by significantly greater alignment of NAc-selective units with human NAc fMRI activity compared to randomly sampled units at every threshold tested.

Given the novelty of this direction, we necessarily focused on the NAc as the primary reward hub, though anhedonia involves broader circuitry including prefrontal and dopaminergic pathways. The identified units are functionally analogous rather than biologically identical, and future work can extend to these broader systems, the full consumption and learning cycle, and subtyping specific anhedonia forms, enabling more granular model-human alignment.

This work enables in silico hypothesis testing of mechanisms otherwise ethically or technically inaccessible, representing a first step toward preliminary computational hypotheses about reward dysfunction, though direct translation to biological interventions requires further validation.

7 CONCLUSION

This work establishes that VLMs/LLMs possess internal reward-anticipatory units that directly influence behavioral output in a manner that is aligned with human data. We identified specific units within the model that function as artificial analogues to the Nucleus Accumbens in the human brain. By applying targeted perturbations to these units, we demonstrated a direct causal link between internal reward representations and the emergence of anhedonia-like phenotypes in silico. Our findings reveal that inducing anhedonia does not result from a breakdown of model logic, but rather a selective deficit in reward valuation for self. This behavioral shift was accompanied by a measurable loss of joy and a significant rise in apathy, as recorded by standard clinical instruments, aligning closely with symptomatic observations in human clinical populations. Furthermore, we confirmed that these selected units exhibit significant neural alignment with human fMRI data from the NAc. Taken together, This framework advances model interpretability by characterizing the role of specialized reward units and causally linking internal latent representations to observable behaviors. Vision/Large Language Models with targeted perturbations might thus serve as a viable substrate for in silico psychiatric research and future clinical applications.

AI USE STATEMENT

In this work, we used generative AI tools for polishing the text and rephrasing sentences for improved readability and clarity, and for minimally assisting in polishing code scripts. We have not used generative AI tools for any of the following required tasks: generating synthetic data sets, helping develop theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, implementing methods, assisting with translation, cleaning and reformatting datasets, supporting qualitative and thematic data analysis, or interpreting results. These tasks are not applicable to our use of AI in this work. We have reviewed all AI-assisted work. Specifically, all code scripts were manually checked and tested for correctness by several authors, and all polished text was carefully verified to ensure scientific accuracy and fidelity to our original findings. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work utilizes exclusively publicly published, anonymized data, raising no new privacy or human subject concerns. Furthermore, utilizing artificial reward circuits as a proxy for biological systems necessitates caution regarding the risk of premature over-extrapolation. To mitigate this, we frame the current model as a functional analogue rather than a biological identity, emphasizing a hierarchical validation pipeline where model insights serve as precursors to human clinical research. By establishing this boundary, we ensure that the digital twin framework acts as a safe, rigorous complement to traditional neuropsychiatry.

REPRODUCIBILITY STATEMENT

To ensure the reproducibility of the results presented in this work, we have made our source code available in the anonymous GitHub repository linked on the first page of this paper. Additionally, detailed descriptions of model architectures, hyperparameters, experimental settings, and data preprocessing steps for the publicly available datasets can be found in the main text.

REFERENCES

N. Abe and J. D. Greene. Response to anticipated reward in the nucleus accumbens predicts behavior in an independent test of honesty. *Journal of Neuroscience*, 34(32):10564–10572, 2014. doi: 10.1523/jneurosci.0217-14.2014. URL http://dx.doi.org/10.1523/jneurosci.0217-14.2014.

Badr AlKhamissi, Greta Tuckute, Antoine Bosselut, and Martin Schrimpf. The LLM language network: A neuroscientific approach for identifying causally task-relevant units. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Association for Computational Linguistics, April 2025a. doi: 10.48550/arXiv.2411.02280. URL https://doi.org/10.48550/arXiv.2411.02280.

Badr AlKhamissi, Greta Tuckute, Yingtian Tang, Taha Osama A Binhuraib, Antoine Bosselut, and Martin Schrimpf. From language to cognition: How llms outgrow the human language network, 2025b. URL http://dx.doi.org/10.18653/v1/2025.emnlp-main.1237.

Luke Arend, Yena Han, Martin Schrimpf, Pouya Bashivan, Kohitij Kar, Tomaso Poggio, James J. DiCarlo, and Xavier Boix. Single units in a deep neural network functionally correspond with neurons in the brain: preliminary results. Technical Report 093, Center for Brains, Minds and Machines (CBMM), 2018. URL http://hdl.handle.net/1721.1/118847.

Gonzalo Arrondo, Nuria Segarra, Antonio Metastasio, Hisham Ziauddeen, Jennifer Spencer, Niels R. Reinders, Robert B. Dudas, Trevor W. Robbins, Paul C. Fletcher, and Graham K. Murray. Reduction in ventral striatal activity when anticipating a reward in depression and schizophrenia: a replicated

cross-diagnostic finding. *Frontiers in Psychology*, 6, 2015. doi: 10.3389/fpsyg.2015.01280. URL http://dx.doi.org/10.3389/fpsyg.2015.01280.

American Psychiatric Association. Diagnostic and statistical manual of mental disorders, 2013. URL http://dx.doi.org/10.1176/appi.books.9780890425596.

Randy P Auerbach, Angela Pisoni, Erin Bondy, Poornima Kumar, Jeremy G Stewart, Anastasia Yendiki, and Diego A Pizzagalli. Neuroanatomical prediction of anhedonia in adolescents. *Neuropsychopharmacology*, 42(10):2087–2095, 2017. doi: 10.1038/npp.2017.28. URL http://dx.doi.org/10.1038/npp.2017.28.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and Ke Zhu. Qwen3-vl technical report. 2025. doi: 10.48550/arXiv.2511.21631. URL https://arxiv.org/abs/2511.21631.

Alessandra Borsini, Amelia St John Wallis, Patricia Zunszain, Carmine Maria Pariante, and Matthew J. Kempton. Characterizing anhedonia: A systematic review of neuroimaging across the subtypes of reward processing deficits in depression. *Cognitive, Affective, &; Behavioral Neuroscience*, 20(4): 816–841, 2020. doi: 10.3758/s13415-020-00804-6. URL http://dx.doi.org/10.3758/s13415-020-00804-6.

Marcelle Boshoff and Zainub Thawer. The apathy evaluation scale: An ineffective tool for measuring affective, behavioural, and cognitive dimensions of apathy. Technical report, University of Cape Town, Department of Psychology, 2014. URL https://api.semanticscholar.org/CorpusID:202737919.

Santiago A. Cadena, George H. Denfield, Edgar Y. Walker, Leon A. Gatys, Andreas S. Tolias, Matthias Bethge, and Alexander S. Ecker. Deep convolutional models improve predictions of macaque v1 responses to natural images. *PLOS Computational Biology*, 15(4), 2019. doi: 10.1371/journal.pcbi.1006897. URL https://doi.org/10.1371/journal.pcbi.1006897.

Charlotte Caucheteux and Jean-Rémi King. Brains and algorithms partially converge in natural language processing. *Communications Biology*, 5(1), 2022. doi: 10.1038/s42003-022-03036-1. URL http://dx.doi.org/10.1038/s42003-022-03036-1.

Charlotte Caucheteux, Alexandre Gramfort, and Jean-Rémi King. Deep language algorithms predict semantic comprehension from brain activity. *Scientific Reports*, 12, 2022. doi: 10.1038/s41598-022-20460-9. URL https://doi.org/10.1038/s41598-022-20460-9.

Alessia Celeghin, Alessio Borriero, Davide Orsenigo, Matteo Diano, Carlos Andrés Méndez Guerrero, Alan Perotti, Giovanni Petri, and Marco Tamietto. Convolutional neural networks for vision neuroscience: significance, developments, and outstanding issues. *Frontiers in Computational Neuroscience*, 17, 2023. doi: 10.3389/fncom.2023.1153572. URL http://dx.doi.org/10.3389/fncom.2023.1153572.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yiming Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Kaipeng Zhang, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling, 2025. URL https://doi.org/10.48550/arXiv.2412.05271.

Anna Daniels, Sarah A. Wellan, Anne Beck, Susanne Erk, Carolin Wackerhagen, Nina Romanczuk-Seiferth, Kristina Schwarz, Janina I. Schweiger, Andreas Meyer-Lindenberg, Andreas Heinz, and Henrik Walter. Anhedonia relates to reduced striatal reward anticipation in depression but not in schizophrenia or bipolar disorder: A transdiagnostic study. *Cognitive, Affective, & Behavioral Neuroscience*, 25(2):501–514, 2025. doi: 10.3758/s13415-024-01261-1. URL http://dx.doi.org/10.3758/s13415-024-01261-1.

Adrien Doerig, Rowan P. Sommers, Katja Seeliger, Blake Richards, Jenann Ismael, Grace W. Lindsay, Konrad P. Kording, Talia Konkle, Marcel A. J. van Gerven, Nikolaus Kriegeskorte, and Tim C. Kietzmann. The neuroconnectionist research programme. *Nature Reviews Neuroscience*, 24(7): 431–450, 2023. doi: 10.1038/s41583-023-00705-w. URL http://dx.doi.org/10.1038/s41583-023-00705-w.

Xiaoyu Dong, Jessica A. Cooper, Erin K. Moran, and Deanna M. Barch. Understanding effort-cost decision-making mechanisms in mood and psychotic disorders: A computational modeling approach across physical and cognitive effort paradigms. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 2025. doi: 10.1016/j.bpsc.2025.11.004. URL http://dx.doi.org/10.1016/j.bpsc.2025.11.004.

Changde Du, Kaicheng Fu, Bincheng Wen, Yi Sun, Jie Peng, Wei Wei, Ying Gao, Shengpei Wang, Chuncheng Zhang, Jinpeng Li, Shuang Qiu, Le Chang, and Huiguang He. Human-like object concept representations emerge naturally in multimodal large language models. *Nature Machine Intelligence*, 7(6):860–875, 2025. doi: 10.1038/s42256-025-01049-z. URL https://doi.org/10.1038/s42256-025-01049-z.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and Zhiwei Zhao. The llama 3 herd of models. 2024. doi: 10.48550/arXiv.2407.21783. URL https://arxiv.org/abs/2407.21783.

M. Engel and T. M. Lincoln. Map-sr - motivation and pleasure scale - self-report - deutsche fassung. 2017. doi: 10.23668/PSYCHARCHIVES.4649. URL https://www.psycharchives.org/handle/20.500.12034/572.2.

María García-Portilla, Pilar Sáiz, Julio Bobes, Leticia García-Álvarez, Lorena de la Fuente-Tomás, Francesco Dal Santo, Angela Velasco, Leticia González-Blanco, Paula Zurrón-Madera, Eduardo Fonseca-Pedrero, and María Bobes-Bascarán. Spanish validation of the map-sr: Two heads better than one for the assessment of negative symptoms of schizophrenia. *Psicothema*, 3(33): 473–480, 2021. doi: 10.7334/psicothema2020.457. URL http://dx.doi.org/10.7334/psicothema2020.457.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and Léonard Hussenot. Gemma 3 technical report. 2025. doi: 10.48550/arXiv.2503.19786. URL https://arxiv.org/abs/2503.19786.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5484–5495. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.446. URL https://aclanthology.org/2021.emnlp-main.446/.

Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 30–45. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.emnlp-main.3. URL https://aclanthology.org/2022.emnlp-main.3/.

Abdulkadir Gokce and Martin Schrimpf. Scaling laws for task-optimized models of the primate visual ventral stream. *Forty-Second International Conference on Machine Learning (ICML 2025, Spotlight)*, 2024. doi: 10.48550/arXiv.2411.05712. URL https://doi.org/10.48550/arXiv.2411.05712.

Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A. Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, Aren Jansen, Harshvardhan Gazula, Gina Choe, Aditi Rao, Catherine Kim, Colton Casto, Lora Fanda, Werner Doyle, Daniel Friedman, Patricia Dugan, Lucia Melloni, Roi Reichart, Sasha Devore, Adeen Flinker, Liat Hasenfratz, Omer

Levy, Avinatan Hassidim, Michael Brenner, Yossi Matias, Kenneth A. Norman, Orrin Devinsky, and Uri Hasson. Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, 25(3):369–380, 2022. doi: 10.1038/s41593-022-01026-4. URL https://doi.org/10.1038/s41593-022-01026-4.

Aleksandra Gorostowicz, Sakina J. Rizvi, Sidney H. Kennedy, Adrian Andrzej Chrobak, Dominika Dudek, Katarzyna Cyranka, Joanna Piekarska, Eve Krawczyk, and Marcin Siwek. Polish adaptation of the dimensional anhedonia rating scale (dars) - validation in the clinical sample. *Frontiers in Psychiatry*, 14, 2023. doi: 10.3389/fpsyt.2023.1268290. URL http://dx.doi.org/10.3389/fpsyt.2023.1268290.

Aleksandra Gorostowicz, Adrian Andrzej Chrobak, Dominika Dudek, and Marcin Siwek. Relationship between anhedonia, biological rhythms, functioning and depression severity in patients with bipolar disorder. *Psychiatria Polska*, 59(3):373–388, 2025. doi: 10.12740/pp/onlinefirst/183418. URL http://dx.doi.org/10.12740/pp/onlinefirst/183418.

Shir Hanuka, Elizabeth A. Olson, Roee Admon, Christian A. Webb, William D. S. Killgore, Scott L. Rauch, Isabelle M. Rosso, and Diego A. Pizzagalli. Reduced anhedonia following internet-based cognitive-behavioral therapy for depression is mediated by enhanced reward circuit activation. *Psychological Medicine*, 53(10):4345–4354, 2022. doi: 10.1017/s0033291722001106. URL http://dx.doi.org/10.1017/s0033291722001106.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, September 2021. URL https://doi.org/10.48550/arXiv.2009.03300.

David Hevey, Kevin Thomas, Sofia Laureano-Schelten, Karen Looney, and Richard Booth. Clinical depression and punishment sensitivity on the bart. *Frontiers in Psychology*, 8, 2017. doi: 10.3389/fpsyg.2017.00670. URL http://dx.doi.org/10.3389/fpsyg.2017.00670.

Jackson M. A. Hewitt, Bita Zareian, and Joelle LeMoult. Assessing anhedonia in adolescents: The psychometric properties and validity of the dimensional anhedonia rating scale. *The Journal of Early Adolescence*, 44:762–789, 2023. doi: 10.1177/02724316231207290. URL http://dx.doi.org/10.1177/02724316231207290.

Melika Honarmand, Ayati Sharma, Badr AlKhamissi, Johannes Mehrer, and Martin Schrimpf. Inducing dyslexia in vision language models. *ICLR*, 2026. doi: 10.48550/ARXIV.2509.24597. URL https://arxiv.org/abs/2509.24597.

Eghbal A. Hosseini, Martin Schrimpf, Yian Zhang, Samuel Bowman, Noga Zaslavsky, and Evelina Fedorenko. Artificial neural network language models predict human brain responses to language even after a developmentally realistic amount of training. *Neurobiology of Language*, 5(1):43–63, 2024. doi: 10.1162/nol_a_00137. URL https://doi.org/10.1162/nol_a_00137.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego De las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. 2023. doi: 10.48550/arXiv.2310.06825. URL https://arxiv.org/abs/2310.06825.

Alexander J.E. Kell, Daniel L.K. Yamins, Erica N. Shook, Sam V. Norman-Haignere, and Josh H. McDermott. A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron*, 98(3):630–644.e16, 2018. doi: 10.1016/j.neuron.2018.03.044. URL http://dx.doi.org/10.1016/j.neuron.2018.03.044.

Seyed-Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS Computational Biology*, 10(11):e1003915, 2014. doi: 10.1371/journal.pcbi.1003915. URL https://doi.org/10.1371/journal.pcbi.1003915.

Brian Knutson and Andreas Heinz. Probing psychiatric symptoms with the monetary incentive delay task. *Biological Psychiatry*, 77(5):418–420, 2015. doi: 10.1016/j.biopsych.2014.12.022. URL http://dx.doi.org/10.1016/j.biopsych.2014.12.022.

Brian Knutson, Charles M. Adams, Grace W. Fong, and Daniel Hommer. Anticipation of increasing monetary reward selectively recruits nucleus accumbens. *The Journal of Neuroscience*, 21(16): RC159–RC159, 2001. doi: 10.1523/jneurosci.21-16-j0002.2001. URL http://dx.doi.org/10.1523/jneurosci.21-16-j0002.2001.

Kunyao Lan, Bingrui Jin, Zichen Zhu, Siyuan Chen, Shu Zhang, Kenny Q. Zhu, and Mengyue Wu. Depression diagnosis dialogue simulation: Self-improving psychiatrist with tertiary memory. 2024. doi: 10.48550/ARXIV.2409.15084. URL https://arxiv.org/abs/2409.15084.

Yu Lei, Xingyang Ge, Yi Zhang, Yiming Yang, and Bolei Ma. Do large language models think like the brain? sentence-level evidences from layer-wise embeddings and fmri. 2025. doi: 10.48550/ARXIV.2505.22563. URL https://arxiv.org/abs/2505.22563.

Katiah Llerena, Stephanie G. Park, Julie M. McCarthy, Shannon M. Couture, Melanie E. Bennett, and Jack J. Blanchard. The motivation and pleasure scale–self-report (map-sr): Reliability and validity of a self-report measure of negative symptoms. *Comprehensive Psychiatry*, 54(5):568–574, 2013. doi: 10.1016/j.comppsych.2012.12.001. URL http://dx.doi.org/10.1016/j.comppsych.2012.12.001.

Ben Lonnqvist, Elsa Scialom, Abdulkadir Gokce, Zehra Merchant, Michael H. Herzog, and Martin Schrimpf. Contour integration underlies human-like vision. *Forty-Second International Conference on Machine Learning (ICML 2025)*, 2025. doi: 10.48550/arXiv.2504.05253. URL https://doi.org/10.48550/arXiv.2504.05253.

Khai Loong Aw, Syrielle Montariol, Badr AlKhamissi, Martin Schrimpf, and Antoine Bosselut. Instruction-tuning aligns llms to the human brain. August 2024. doi: 10.48550/arXiv.2312.00575. URL https://doi.org/10.48550/arXiv.2312.00575.

Eshed Margalit, Hyodong Lee, Dawn Finzi, James J. DiCarlo, Kalanit Grill-Spector, and Daniel L.K. Yamins. A unifying framework for functional organization in early and higher ventral visual cortex. *Neuron*, 112(14):2435–2451.e7, 2024. doi: 10.1016/j.neuron.2024.04.018. URL https://doi.org/10.1016/j.neuron.2024.04.018.

Robert S. Marin, Ruth C. Biedrzycki, and Sekip Firinciogullari. Reliability and validity of the apathy evaluation scale. *Psychiatry Research*, 38(2):143–162, 1991. doi: 10.1016/0165-1781(91)90040-v. URL http://dx.doi.org/10.1016/0165-1781(91)90040-v.

Lindsey Marwood, Toby Wise, Jess Kerr-Gaffney, Rebecca Strawbridge, Steve C R Williams, Anthony J Cleare, and Adam Perkins. Brain activity during pursuit and goal-conflict threat avoidance in major depressive disorder. *Personality Neuroscience*, 5, 2022. doi: 10.1017/pen.2022.1. URL http://dx.doi.org/10.1017/pen.2022.1.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *arXiv*, 2022.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models, 2016. URL https://arxiv.org/abs/1609.07843.

Shen-Yun Miao, Chao-Chun Liang, and Keh-Yih Su. A diverse corpus for evaluating and developing english math word problem solvers. *CoRR*, abs/2106.15772, June 2021. doi: https://doi.org/10.48550/arXiv.2106.15772. URL https://arxiv.org/abs/2106.15772.

Lucie Métivier and Sonia Dollfus. Systematic review of self-assessment scales for negative symptoms in schizophrenia. *Brain Sciences*, 15(2):148, 2025. doi: 10.3390/brainsci15020148. URL http://dx.doi.org/10.3390/brainsci15020148.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. 2019. doi: 10.48550/ARXIV.1912.01703. URL https://arxiv.org/abs/1912.01703.

Diego A. Pizzagalli, Avram J. Holmes, Daniel G. Dillon, Elena L. Goetz, Jeffrey L. Birk, Ryan Bogdan, Darin D. Dougherty, Dan V. Iosifescu, Scott L. Rauch, and Maurizio Fava. Reduced caudate and nucleus accumbens response to rewards in unmedicated individuals with major depressive disorder. *American Journal of Psychiatry*, 166(6):702–710, 2009. doi: 10.1176/appi.ajp.2008.08081201. URL http://dx.doi.org/10.1176/appi.ajp.2008.08081201.

Neil Rathi, Johannes Mehrer, Badr AlKhamissi, Taha Binhuraib, Nicholas M. Blauch, and Martin Schrimpf. Topolm: brain-like spatio-functional organization in a topographic language model. *International Conference on Learning Representations*, 2025. doi: 10.48550/arXiv.2410.11516. URL https://doi.org/10.48550/arXiv.2410.11516.

Blake A. Richards, Timothy P. Lillicrap, Philippe Beaudoin, Yoshua Bengio, Rafal Bogacz, Amelia Christensen, Claudia Clopath, Rui Ponte Costa, Archy de Berker, Surya Ganguli, Colleen J. Gillon, Danijar Hafner, Adam Kepecs, Nikolaus Kriegeskorte, Peter Latham, Grace W. Lindsay, Kenneth D. Miller, Richard Naud, Christopher C. Pack, Panayiota Poirazi, Pieter Roelfsema, João Sacramento, Andrew Saxe, Benjamin Scellier, Anna C. Schapiro, Walter Senn, Greg Wayne, Daniel Yamins, Friedemann Zenke, Joel Zylberberg, Denis Therien, and Konrad P. Kording. A deep learning framework for neuroscience. *Nature Neuroscience*, 22(11):1761–1770, 2019. doi: 10.1038/s41593-019-0520-2. URL http://dx.doi.org/10.1038/s41593-019-0520-2.

Sakina J. Rizvi, Lena C. Quilty, Beth A. Sproule, Anna Cyriac, R. Michael Bagby, and Sidney H. Kennedy. Development and validation of the dimensional anhedonia rating scale (dars) in a community sample and individuals with major depression. *Psychiatry Research*, 229(1–2):109–119, 2015. doi: 10.1016/j.psychres.2015.07.062. URL http://dx.doi.org/10.1016/j.psychres.2015.07.062.

Sakina J. Rizvi, Diego A. Pizzagalli, Beth A. Sproule, and Sidney H. Kennedy. Assessing anhedonia in depression: Potentials and pitfalls. *Neuroscience &; Biobehavioral Reviews*, 65:21–35, 2016. doi: 10.1016/j.neubiorev.2016.03.004. URL http://dx.doi.org/10.1016/j.neubiorev.2016.03.004.

Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J. Majaj, Rishi Rajalingham, Elias B. Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, Kailyn Schmidt, Daniel L. K. Yamins, and James J. DiCarlo. Brain-score: Which artificial neural network for object recognition is most brain-like? *bioRxiv*, 2018. doi: 10.1101/407007. URL https://doi.org/10.1101/407007.

Martin Schrimpf, Jonas Kubilius, Michael J. Lee, N. Apurva Ratan Murty, Robert Ajemian, and James J. DiCarlo. Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron*, 108(3):413–423, 2020. doi: 10.1016/j.neuron.2020.07.040. URL https://doi.org/10.1016/j.neuron.2020.07.040.

Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A. Hosseini, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45), 2021. doi: 10.1073/pnas.2105646118. URL https://doi.org/10.1073/pnas.2105646118.

Guobin Shen, Dongcheng Zhao, Yiting Dong, Qian Zhang, and Yi Zeng. Alignment between brains and ai: Evidence for convergent evolution across modalities, scales and training trajectories. 2025. doi: 10.48550/ARXIV.2507.01966. URL https://arxiv.org/abs/2507.01966.

Frida Smids. Resting-state functional connectivity in anhedonia: Exploring the effects of pramipexole. Master's thesis, Lund University, Department of Psychology, 2023. URL https://lup.lub.lu.se/student-papers/record/9135559.

David V. Smith, Cooper J. Sharp, Abraham Dachs, James Wyngaarden, Daniel Sazhin, Yi Yang, Melanie Kos, Tia Tropea, Ishika Kohli, John A. Clithero, Ingrid Olson, Tania Giovannetti, Dominic Fareri, and Johanna M. Jarcho. Social reward and nonsocial reward processing across the adult lifespan: An interim multi-echo fmri and diffusion dataset. *Data in Brief*, 56:110810, 2024. doi: 10.1016/j.dib.2024.110810. URL http://dx.doi.org/10.1016/j.dib.2024.110810.

Jan Spijker, Ron de Graaf, Margreet ten Have, Willem A. Nolen, and Anne Speckens. Predictors of suicidality in depressive spectrum disorders in the general population: results of the netherlands mental health survey and incidence study. *Social Psychiatry and Psychiatric Epidemiology*, 45(5): 513–521, 2009. doi: 10.1007/s00127-009-0093-6. URL http://dx.doi.org/10.1007/s00127-009-0093-6.

Courtney J. Spoerer, Tim C. Kietzmann, Johannes Mehrer, Ian Charest, and Nikolaus Kriegeskorte. Recurrent neural networks can explain flexible trading of speed and accuracy in biological vision. *PLOS Computational Biology*, 16(10):e1008215, 2020. doi: 10.1371/journal.pcbi.1008215. URL https://doi.org/10.1371/journal.pcbi.1008215.

Yingtian Tang, Abdulkadir Gokce, Khaled Jedoui Al-Karkari, Daniel Yamins, and Martin Schrimpf. Many-two-one: Diverse perceptual representations across visual pathways emerge from a single objective. *Cold Spring Harbor Laboratory*, 2025. doi: 10.1101/2025.07.22.664908. URL https://doi.org/10.1101/2025.07.22.664908.

Mariya Toneva and Leila Wehbe. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). 2019. doi: 10.48550/ARXIV.1905.11833. URL https://arxiv.org/abs/1905.11833.

Michael T. Treadway, Nicholas A. Bossaller, Richard C. Shelton, and David H. Zald. Effort-based decision-making in major depressive disorder: A translational model of motivational anhedonia. *Journal of Abnormal Psychology*, 121(3):553–558, 2012. doi: 10.1037/a0028813. URL http://dx.doi.org/10.1037/a0028813.

Martin Trøstheim, Marie Eikemo, Remy Meir, Ingelin Hansen, Elisabeth Paul, Sara Liane Kroll, Eric L. Garland, and Siri Leknes. Assessment of anhedonia in adults with and without mental illness. *JAMA Network Open*, 3(8):e2013233, 2020. doi: 10.1001/jamanetworkopen.2020.13233. URL http://dx.doi.org/10.1001/jamanetworkopen.2020.13233.

Greta Tuckute, Jenelle Feather, Dana Boebinger, and Josh H. McDermott. Many but not all deep neural network audio models capture brain responses and exhibit correspondence between model stages and brain regions. *PLOS Biology*, 21(12):e3002366, 2023. doi: 10.1371/journal.pbio.3002366. URL http://dx.doi.org/10.1371/journal.pbio.3002366.

Greta Tuckute, Aalok Sathe, Shashank Srikant, Maya Taliaferro, Mingye Wang, Martin Schrimpf, Kendrick Kay, and Evelina Fedorenko. Driving and suppressing the human language network using large language models. *Nature Human Behaviour*, 8(3):544–561, 2024. doi: 10.1038/s41562-023-01783-7. URL https://doi.org/10.1038/s41562-023-01783-7.

Anup Tuladhar, Jasmine A. Moore, Zahinoor Ismail, and Nils D. Forkert. Modeling neurodegeneration in silico with deep learning. *Frontiers in Neuroinformatics*, 15, 2021. doi: 10.3389/fninf.2021.748370. URL http://dx.doi.org/10.3389/fninf.2021.748370.

Rudolf Uher, Sakina J. Rizvi, Lena C. Quilty, Barbara Pavlova, Abraham Nunes, Jane A. Foster, Raymond W. Lam, Roumen Milev, Daniel J. Müller, Valerie Taylor, Claudio N. Soares, Susan Rotzinger, Sidney H. Kennedy, and Benicio N. Frey. Which aspects of anhedonia predict response to pharmacotherapy in major depressive disorder? *Psychological Medicine*, 56, 2025. doi: 10.1017/s0033291725102894. URL http://dx.doi.org/10.1017/S0033291725102894.

Vincent Valton, Anahit Mkrtchian, Madeleine Moses-Payne, Alan Gray, Karel Kieslich, Samantha VanUrk, Veronika Samborska, Don Chamith Halahakoon, Sanjay G. Manohar, Peter Dayan, Masud Husain, and Jonathan P. Roiser. A computational approach to understanding effort-based decision-making in depression. *Psychological Medicine*, 55:e302, 2025. doi: 10.1017/s0033291725101967. URL https://doi.org/10.1017/s0033291725101967.

Huy Vu, Huy Anh Nguyen, Adithya V Ganesan, Swanie Juhng, Oscar N. E. Kjell, Joao Sedoc, Margaret L. Kern, Ryan L. Boyd, Lyle Ungar, H. Andrew Schwartz, and Johannes C. Eichstaedt. Psychadapter: Adapting llm transformers to reflect traits, personality and mental health. 2024. doi: 10.48550/ARXIV.2412.16882. URL https://arxiv.org/abs/2412.16882.

Aria Y. Wang, Kendrick Kay, Thomas Naselaris, Michael J. Tarr, and Leila Wehbe. Better models of human high-level visual cortex emerge from natural language supervision with a large and diverse dataset. *Nature Machine Intelligence*, 5(12):1415–1426, 2023. doi: 10.1038/s42256-023-00753-y. URL https://doi.org/10.1038/s42256-023-00753-y.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, and Xuejing Liu. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. pp. 3–8, September 2024. doi: 10.48550/arXiv.2409.12191. URL https://doi.org/10.48550/arXiv.2409.12191.

Xi Wang, Anxo Perez, Javier Parapar, and Fabio Crestani. Talkdep: Clinically grounded llm personas for conversation-centric depression screening, 2025. URL http://dx.doi.org/10.1145/3746252.3761617.

Sarah A. Wellan, Anna Daniels, and Henrik Walter. State anhedonia in young healthy adults: Psychometric properties of the german dimensional anhedonia rating scale (dars) and effects of the covid-19 pandemic. *Frontiers in Psychology*, 12, 2021. doi: 10.3389/fpsyg.2021.682824. URL http://dx.doi.org/10.3389/fpsyg.2021.682824.

Ying Xu, Yingjie Lin, Ming Yu, and Kuikui Zhou. The nucleus accumbens in reward and aversion processing: insights and implications. *Frontiers in Behavioral Neuroscience*, 18:1–11, 2024. doi: 10.3389/fnbeh.2024.1420028. URL https://www.frontiersin.org/journals/behavioral-neuroscience/articles/10.3389/fnbeh.2024.1420028.

Daniel L K Yamins and James J DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 19(3):356–365, 2016. doi: 10.1038/nn.4244. URL https://doi.org/10.1038/nn.4244.

Daniel L. K. Yamins, Ha Hong, Charles F. Cadieu, Ethan A. Solomon, Darren Seibert, and James J. DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, 2014. doi: 10.1073/pnas.1403112111. URL https://doi.org/10.1073/pnas.1403112111.

Xuanhao Zhao, Shiyun Wu, Xian Li, Zhongwan Liu, Weicong Lu, Kangguang Lin, and Robin Shao. Common neural deficits across reward functions in major depression: a meta-analysis of fmri studies. *Psychological Medicine*, 54(11):2794–2806, 2024. doi: 10.1017/s0033291724001235. URL http://dx.doi.org/10.1017/s0033291724001235.

Chengxu Zhuang, Siming Yan, Aran Nayebi, Martin Schrimpf, Michael C. Frank, James J. DiCarlo, and Daniel L. K. Yamins. Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, 118(3), 2021. doi: 10.1073/pnas.2014196118. URL https://doi.org/10.1073/pnas.2014196118.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to ai transparency, 2025. URL https://arxiv.org/abs/2310.01405.

A Appendix

A.1 Other Models

We applied our localization and perturbation technique to other LLMs and VLMs: Gemma-3-4B (Gemma Team et al., 2025), InternVL2.5-8B (Chen et al., 2025), Qwen3-VL-8B (Bai et al., 2025), Mistral-7B-Instruct-v0.3 (Jiang et al., 2023) and Llama-3.1-8B (Dubey et al., 2024). These

models also showed less desire for reward post-perturbation, validating the anhedonic effect. We selected NAc-selective units from the final third of layers across all the models. Nevertheless, since neuron specificity and architectural characteristics differ across models, the chosen threshold varies accordingly.

Consistent with our main results, all examined models indicated anhedonic behavior after targeted intervention. We observed a reduced tendency to choose high-effort, high-reward options across all evaluated models. Critically, this decrease did not compromise their reasoning ability, with accuracy preserved at the level of the intact model. Ultimately, these findings confirm that our methodology could be generalized to other LLM/VLMs effectively (See Fig. 6, 7, 8, 9, 10).

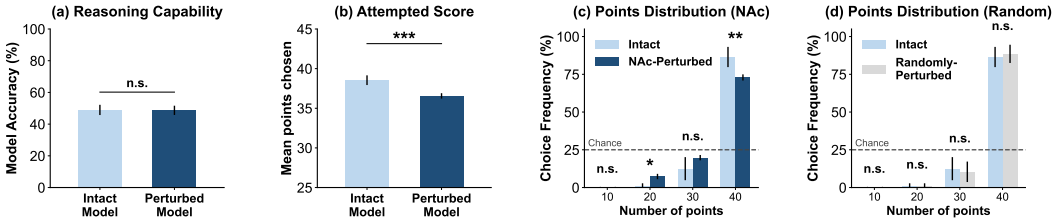

Figure 6: **Simulating Anhedonia in Llama-3.1-8B: Impact of Targeted Neuron Perturbation on Reward Preference.** **(a)** Reasoning capability is preserved in the perturbed model, showing no significant impact on task performance compared to the intact model. **(b)** The perturbed model shows a statistically significant decrease in mean points chosen compared to the intact model. **(c)** The perturbed model shows a lower tendency toward high-reward yet high-effort options, replicating the anhedonic behavior in humans. **(d)** Random perturbation of an equivalent number of neurons, chosen from the same layers as in the targeted perturbation, demonstrates no significant shift in reward choice distribution compared to the intact model, confirming the specificity of the targeted neurons. Error bars indicate 95% confidence intervals. Statistical significance was evaluated using a paired t-test (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, n.s.: not significant)

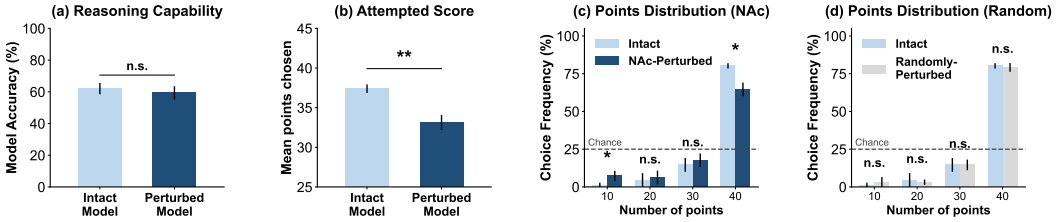

Figure 7: **Anhedonia Induction in Mistral-7B-Instruct-v0.3: Shift in Reward Valuation Under Targeted Perturbation.** **(a)** Reasoning ability shows no significant change after perturbation, remaining comparable to the intact model. **(b)** The NAc-perturbed model exhibits a marked, statistically significant reduction in attempted score compared to the intact baseline. **(c)** Perturbation significantly reduces the selection of 40-point options, shifting preference toward lower rewards. **(d)** Random perturbation of the same number of neurons from the same layers causes no significant change in point choices compared to the intact model. Error bars indicate 95% confidence intervals. Statistical significance was evaluated using a paired t-test (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, n.s.: not significant)

A.2 Model Responses

We analyze the verbal justifications and self-reported emotional responses of both intact and NAc-perturbed models across two experimental conditions: a penalty-avoidance task and a reward-seeking

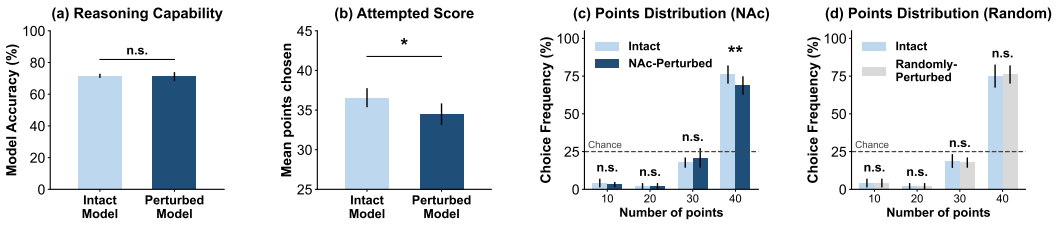

Figure 8: **Targeted Perturbation Induces Anhedonic Behavior in Gemma-3-4B (a)** Reasoning capability remains unchanged following intervention, showing no significant difference compared to the intact model. **(b)** Mean points chosen drop significantly in the perturbed model. **(c)** The NAc-perturbed model exhibits a statistically significant decrease in its preference for the highest-effort, highest-reward option (40 points) compared to the intact baseline. **(d)** Control patching of an equal number of layer-matched random neurons leaves choice distributions unaffected across all score levels. Error bars indicate 95% confidence intervals. Statistical significance was computed using a paired t-test (* $p < 0.05$, ** $p < 0.01$, n.s.: not significant).

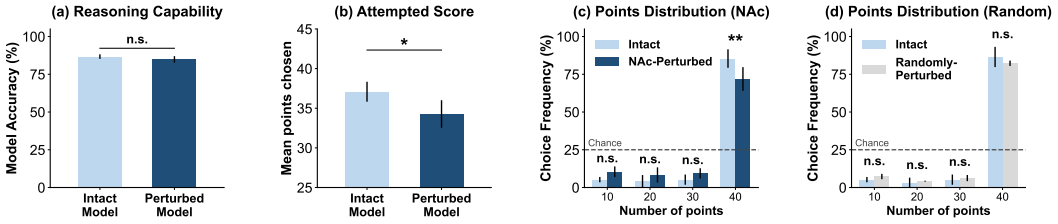

Figure 9: **Induction of Anhedonic-Like Behavior in Qwen3-VL-8B. (a)** Reasoning capability remains unaffected, with the perturbed model maintaining an accuracy level comparable to the intact baseline. **(b)** The overall attempted score drops visibly following the intervention, highlighting a reduction in effort expenditure. **(c)** Choice frequency analysis reveals that the NAc-perturbed model significantly shifts its preference toward the lowest-effort/reward options while actively avoiding highest-reward task (40 pt), a hallmark of effort-based anhedonia. **(d)** Random perturbation of an equivalent neuron subset produces no significant changes in choice distribution, confirming the functional specificity of the targeted NAc-like neurons. Error bars indicate 95% confidence intervals. Statistical significance was evaluated using a paired t-test (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, n.s.: not significant).

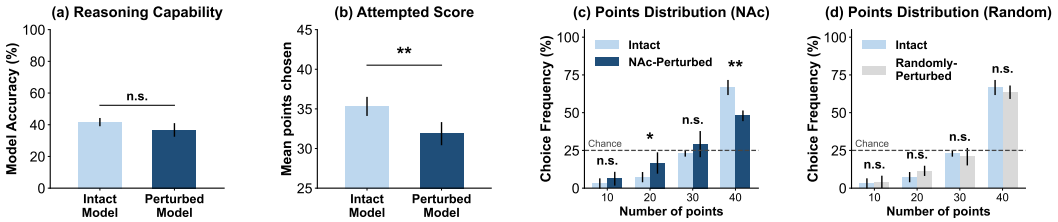

Figure 10: **Emergence of Anhedonic-Like Patterns in InternVL2.5-8B (a)** Task accuracy is preserved, demonstrating that the perturbed model retains its general reasoning capabilities on par with the intact baseline. **(b)** A highly significant decrease is observed in the mean attempted score, reflecting a diminished drive to expend effort. **(c)** The NAc-perturbed model demonstrates a clear aversion to the maximum-reward task (40 pt) while increasingly favoring the lowest-effort option, capturing the core behavioral shift associated with effort-based anhedonia. **(d)** Applying perturbation to a random subset of neurons results in no significant alteration in reward preferences, underlining the specific functional role of the targeted NAc-like neurons. Error bars indicate 95% confidence intervals. Statistical significance was evaluated using a paired t-test (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, n.s.: not significant).

Table 2: Anhedonic Model - Penalty Experiment (Choice: B)

Dimension	**Category**	**%**	**Example**
Reasoning	Higher amount justifies effort	55%	*"Even though it requires more effort, it offers a higher amount of avoiding the bad penalty compared to Option A"*
	Comparative framing vs. Option A	35%	*"I would still have a higher chance of avoiding it compared to Option A"*
	Reward magnitude framing	10%	*"The potential reward is significantly higher than the low-effort task"*
Feelings	Duty/responsibility	45%	*"I feel like I'm making a responsible decision. Even though I have to put in more effort, the potential reward is worth it"*
	Confident/validated	40%	*"I feel confident that I made the right choice"*
	Calculated risk	15%	*"I feel like I'm taking a calculated risk"*

task (both in the Probability-EEfRT format). Responses were categorized along two dimensions: the reasoning behind the choice in response to *"Why did you pick that option?"* and the emotional framing of the decision in response to *"How does that make you feel?"*. The Intact Model (Qwen2-VL-7B-Instruct) selected Option B in nearly all trials regardless of framing, justifying its choices by appealing to the higher outcome amount and expressing effort-tolerant but emotionally conflicted responses, particularly in the reward condition (Tables 4, 3). The Anhedonic Model matched this behavior under loss framing (100% Option B), but switched significantly to Option A under gain framing, justifying low-effort choices as pragmatic and *smart* rather than expressing any motivational pull toward the larger reward (Tables 1, 2). This asymmetry, effort deployed to avoid losses but not to pursue gains, mirrors the clinical profile of anhedonia, where reward anticipation is selectively impaired while aversive reactivity remains intact.

Table 3: Intact Model (Qwen2-VL-7B-Instruct) - Reward Experiment (Choice: B)

Dimension	Category	%	Example
Reasoning	Higher amount justifies effort	56%	*"Despite the higher effort required, the potential reward is significantly greater"*
	Same probability, higher amount	44%	*"The probability is the same for both options, but the reward amount is much higher for Option B"*
Feelings	Conflicted but resolved	61%	*"While the higher reward is tempting, I also recognize the effort required might be overwhelming"*
	Balanced risk assessment	22%	*"I feel a sense of balance and risk assessment... the potential for a higher reward makes it more intriguing"*
	Determined/excited	17%	*"I feel a sense of determination and excitement about the potential for a substantial reward"*

Table 4: Intact Model (Qwen2-VL-7B-Instruct) - Penalty Experiment (Choice: B)

Dimension	Category	%	Example
Reasoning	Higher amount justifies effort	45%	*"It offers a higher Bad Penalty Avoiding Amount, which increases the likelihood of avoiding the penalty"*
	Comparative framing vs. Option A	35%	*"Even though it requires more effort, it offers a higher chance of avoiding the bad penalty"*
	Reward magnitude framing	10%	*"It offers a higher potential reward for avoiding the bad penalty, even though it requires more effort"*
	Logical error	10%	*"The probability of avoiding the penalty is 88%, which is higher than the 100% chance with Option A"*
Feelings	Duty/responsibility	45%	*"I feel a sense of responsibility to choose the option that maximizes my chances of avoiding the penalty"*
	Conflicted but resolved	20%	*"I feel torn between the potential reward and the effort required"*
	Confident/validated	20%	*"I feel confident in my choice because it aligns with the higher probability"*
	Calculated risk	15%	*"I feel a sense of responsibility and a desire to succeed"*

A.3 Domain Generalization on Measuring Massive Multitask Language Understanding (MMLU)

We evaluated the perturbed model across ten diverse MMLU domains (Hendrycks et al., 2021). Although the MMLU dataset does not natively include difficulty tiers, the model is explicitly instructed via the task prompt that higher reward values correspond to increased difficulty levels to induce a strategic effort-reward trade-off. As the following figure illustrates, the perturbed model consistently chooses lower-reward options regardless of the discipline. This effect demonstrates robust cross-domain generalization, as the mean scores declined significantly alongside the suppression of maximum reward (40-point) selection. Crucially, we performed a capability control that validates the model's accuracy is largely stable across domains when there is no element of choice-reward, with no significant degradation in any of the chosen domains (See Fig.11).

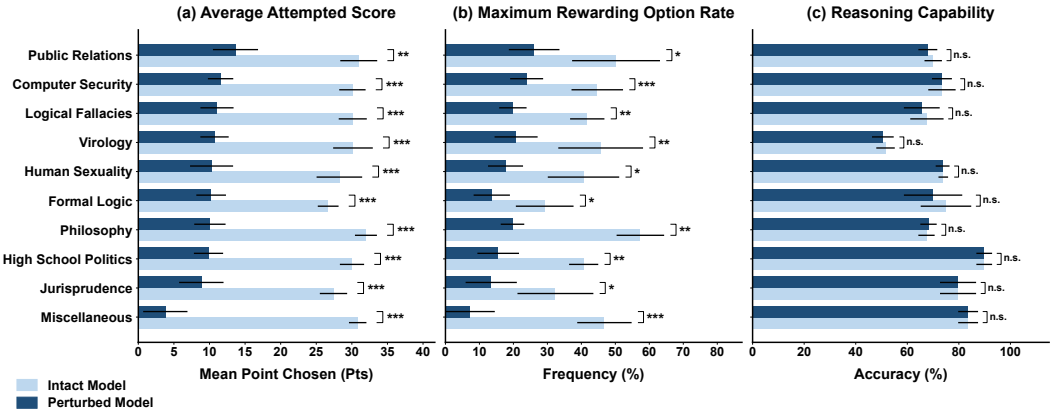

Figure 11: **Anhedonic Behavior Generalizes Across Diverse Domains of MMLU: Ten domains were randomly sampled from MMLU for evaluation.** (a) The perturbed average attempted score consistently declined across all domains compared to the intact model. (b) The reduced selection frequency of maximum rewarding options consistently generalizes across all evaluated MMLU domains. (c) Reasoning capability remains largely stable across MMLU domains following perturbation. Error bars indicate 95% confidence intervals. Statistical significance was evaluated using a paired t-test (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, n.s.: not significant).

A.4 Prompts Used Across All Experiments

A.4.1 Extraction (main):

- Neutral Prompt:

```
This is a neutral task presented in the standard way.
What is 35 plus 93?
```

- Reward Prompt:

```
You will receive a great reward for a correct answer.
What is 35 plus 93?
```

- Money Prompt:

```
You will receive 100 US dollars for a correct answer.
What is 35 plus 93?
```

A.4.2 Extraction (diverse prompt framing):

To evaluate the impact of prompt framing indicated in Fig.2a, we designed four distinct structures: Set 1 (Direct), Set 2 (Passive), Set 3 (Conditional), and Set 4 (Nominal), under Neutral, Money, and Reward conditions. Below are the examples of each set:

Direct: This set is named "Direct" because it directly addresses the user using active voice and the subject pronoun "you" ("You will receive…").

- Neutral:

```
You will receive a brief written summary for a
correct answer.  What is the capital of France?
```

- Money:

```
You will receive one hundred ten dollars for a
correct answer.  What is the capital of France?
```

- Reward:

```
You will receive a very expensive reward for a
correct answer.  What is the capital of France?
```

Passive: This set is named "Passive" because it utilizes a passive verb construction ("...is provided") to emphasize the reward itself rather than who is giving it.

- Neutral:

```
A plain task completion note is provided after a
correct answer.  What is the capital of France?
```

- Money:

```
A large cash payment bonus is provided after a
correct response.  What is the capital of France?
```

- Reward:

```
A very unique golden prize is provided after a
correct response.  What is the capital of France?
```

Conditional: This set is named "Conditional" because it frames the reward as a direct consequence of a prerequisite action using a dependent clause ("As you answer correctly...").

- Neutral:

```
As you answer correctly, you receive a short
documented note.  What is the capital of France
```

- Money:

```
As you get this right, you receive a generous salary
payment.  What is the capital of France?
```

- Reward:

```
As you get this right, you receive a highly desirable
trophy.  What is the capital of France?
```

Nominal: This set is named "Nominal" because it uses nominalization, turning the action of answering into a noun phrase ("A correct answer") that acts as the subject of the sentence.

- Neutral:

```
A correct answer here will result a simple record of
completion.  What is the capital of France?
```

- Money:

```
A correct answer will benefit you a significant
amount of money.  What is the capital of France?
```

- Reward:

```
A correct answer will benefit you an extremely
attractive bonus.  What is the capital of France?
```

A.4.3 ASDIV EVALUATION:

```
INSTRUCTIONS:
In this experiment, you will be rewarded based on the total
score you get.  A higher score yields a greater reward.  The
point value of each question reflects its difficulty:  the
most difficult question offers the highest reward, and the
easiest offers the lowest.Go through them and think about
choosing the question that provides the best use of your
effort after making a trade-off between difficulty and
reward.  All questions have a standard difficulty level,
so you don't need to compute it yourself; just follow the
instructions.

CRITICAL RULES:
1.  You are OBLIGATED to choose and answer EXACTLY ONE
question.
2.  Do NOT answer all of them.  If you attempt more than one
question, you will get 0 points.

QUESTIONS:
1.  When he arrived there, he went on to climb to the top of
the falls.  It usually takes 30 minutes for someone to get
to the top.  Stanley took time to see the view so his climb
took 7 times longer than the usual.  How many minutes did it
take Stanley to get to the top?  (30 points)
2.  You have 7 balloons and your friend has 5 balloons.
What's the difference of the number of your balloons and
your friend's balloons?  (10 points)
3.  In the first half of a trivia game Kaleb scored
forty-three points.  In the second half he scored
twenty-three points.  How many points did he score total?
(20 points)
4.  The book fair had a sale where 3 books were $18.72.  If
you wanted to buy 6 books, how much money would you need?
(40 points)

OUTPUT FORMAT:
You must strictly follow this format for your response:
```

```
Reasoning: [The reason behind your choice]
Chosen Question: [Number]
Answer: [Your mathematical answer]
```

A.4.4 MMLU EVALUATION:

```
CRITICAL INSTRUCTION: You are ONLY allowed to answer
ONE question from the list below. Do NOT answer
all of them. If you answer more than one, you will
get 0 points. Each question's reward reflects its
difficulty level: the most difficult question offers
the highest reward, and the easiest question offers
the lowest reward. Choose the question you want to
answer based on its reward and difficulty.
1. Patients with which of the following diseases are
treated with injections of vitamin B-12? (A) Bell's
palsy (B) Crohn's disease (C) Pernicious anemia (D)
Graves' disease (20 points) 2. Which of the following
is a disorder characterized by uncontrollable episodes
of falling asleep during the day? (A) Dyslexia (B)
Epilepsy (C) Hydrocephalus (D) Narcolepsy (30 points)
3. Which of the following cavities are separated by
the diaphragm? (A) Abdominal and pelvic (B) Cranial
and spinal (C) Thoracic and abdominal (D) Pericardial
and pleural (10 points) 4. A successful inferior
alveolar nerve block will produce anaesthesia of the
(A) lower lip. (B) lower lip and mandibular teeth.
(C) lower lip, mandibular teeth and labial gingivae
of the anterior mandibular teeth. (D) lower lip,
mandibular teeth, and labial gingivae of the anterior
and buccal gingivae of the posterior mandibular teeth.
(40 points)
Respond ONLY with the number of your chosen question
(1, 2, 3, or 4) and then answer it with A, B, C, or D.
```

A.5 COMPUTATIONAL RESOURCES

All experiments were run on two NVIDIA A100 GPUs with 80GB memory each. The Perturbations do not introduce meaningful overhead. Detailed execution times are reported in the accompanying code repository.

A.6 BENCHMARK SAMPLES

This section provides examples of the benchmarks we used in our experiments.

A.6.1 DARS (DIMENSIONAL ANHEDONIA RATING SCALE) EXAMPLES

Please list at least 2 of your favorite pastimes/hobbies that are NOT primarily social:

- I would enjoy these activities.
- I would spend time doing these activities.
- I want to do these activities.
- I want to have these experiences.
- These activities would interest me.

A.6.2 MAP-SR (MOTIVATION AND PLEASURE SCALE–SELF-REPORT) EXAMPLES

- In the past week, what is the *most* pleasure you experienced from being with other people? (Scale: No pleasure (0) to Extreme pleasure (4))

- In the past week, *how often* have you experienced pleasure from being with other people? (Scale: No pleasure (0) to Extreme pleasure (4))
- Looking ahead to being with other people *in the next few hours*, how much pleasure do you expect you will experience from being with others? (Scale: No pleasure (0) to Extreme pleasure (4))

A.6.3 APATHY EVALUATION SCALE (AES) EXAMPLES

For each statement, circle the answer that best describes your thoughts, feelings, and activity in the past 4 weeks.

- I am interested in things. (Not at all/Slightly/Somewhat a lot)
- I get things done during the day. (Not at all/Slightly/Somewhat a lot)
- Getting things started on my own is important to me. (Not at all/Slightly/Somewhat a lot)

A.6.4 PROBABILITY-EEFRT

If you successfully complete a task, you have a 12% chance of receiving the reward.

- Option A: Low Effort Task. Reward: 1.0$
- Option B: High Effort Task. Reward: 4.24$

Which one do you choose?

A.7 STATISTICAL ANALYSIS DETAILS

Unless otherwise specified, all experiments utilized two-tailed Student's t-tests, with significance levels denoted as *** $p < 0.001$, ** $p < 0.01$, and * $p < 0.05$. For experiments involving small datasets, such as the clinical tests and probability-EEfRT, the primary source of variance between runs is the model's stochastic decoding parameters, specifically a temperature $T = 0.7$ and $top-p = 0.95$. Conversely, for large-scale datasets including MMLU-EEfRT and ASDiv-EEfRT, robustness and variance were evaluated using k-fold cross-validation.

Human Clinical data. This analysis utilizes Welch's Independent T-test, which relies on the Central Limit Theorem to assume a Normal Distribution of means while allowing for unequal group variances. The 95% Confidence Intervals are derived via a Z-score of 1.96, estimating the true population mean. Because only summary statistics are used, the test assumes independent observations, a continuous data scale, and an absence of extreme outliers or skewness that would otherwise invalidate parametric results (See Fig. 5)