# Mechanistic Decoding of Cognitive Constructs in LLMs

Yitong Shou, Manhao Guan



***Abstract*—While Large Language Models (LLMs) demonstrate increasingly sophisticated affective capabilities, the internal mechanisms by which they process complex emotions remain unclear. Existing interpretability approaches often treat models as black boxes or focus on coarse-grained basic emotions, leaving the cognitive structure of more complex affective states underexplored. To bridge this gap, we propose a Cognitive Reverse-Engineering framework based on Representation Engineering (RepE) to analyze social-comparison jealousy. By combining appraisal theory with subspace orthogonalization, regression-based weighting, and bidirectional causal steering, we isolate and quantify two psychological antecedents of jealousy, *Superiority of Comparison Person* and *Domain Self-Definitional Relevance*, and examine their causal effects on model judgments. Experiments on eight LLMs from the Llama, Qwen, and Gemma families suggest that models natively encode jealousy as a structured linear combination of these constituent factors. Their internal representations are broadly consistent with the human psychological construct, treating *Superiority* as the foundational trigger and *Relevance* as the ultimate intensity multiplier. Our framework also demonstrates that toxic emotional states can be mechanically detected and surgically suppressed, suggesting a possible route toward representational monitoring and intervention for AI safety in multi-agent environments.**

***Index Terms*—Affective Computing, AI Safety, Mechanistic Interpretability, Representation Engineering, Social-Comparison Jealousy.**

## I. Introduction

The advent of Large Language Models (LLMs) has catalyzed a paradigm shift in Affective Computing (AC). Moving beyond simple sentiment classification, modern LLMs show strong capabilities in affective understanding and generation, supporting applications such as mental health support, companion AI, and interactive non-player characters (NPCs) [1]. As these models are increasingly deployed in socially sensitive contexts, it is important to ensure that their emotional reasoning is transparent, predictable, and aligned with human values.

Despite these advancements, the internal mechanisms by which LLMs process emotions remain unclear. Inference-time methods such as prompt engineering and chain-of-thought prompting rely on textual outputs, but such outputs may constitute unfaithful post-hoc rationalizations. Parameter-adaptation methods such as LoRA fine-tuning improve downstream performance, yet they still treat the model's internal representations as an optimization target rather than an object of cognitive explanation.

Interpretability research offers a possible route forward, but existing paradigms still leave a granularity gap. Traditional bottom-up approaches, such as circuit analysis [2] and sparse autoencoders [3], operate at the level of neurons or local features and are often too microscopic for abstract psychological constructs. Representation Engineering (RepE), by contrast, isolates high-level concepts that emerge from population-level neural activity [4]. Although Tak et al. [5] showed that RepE can probe coarse-grained basic emotions, it remains unclear whether the same paradigm can decode complex emotions whose elicitation depends on specific

This challenge is especially salient for **social-comparison jealousy**. Rather than reflecting general positive or negative valence, jealousy arises from upward social comparison. A mechanistic account therefore requires disentangling the specific antecedents that produce this emotion, not merely relying on general appraisal dimensions. In this study, we focus on two theoretically central antecedents—*Superiority of Comparison Person* and *Domain Self-Definitional Relevance*—and include *Weekday* as a placebo control to test whether the extracted representations reflect meaningful psychological structure.

To address this challenge, we propose a **Cognitive Reverse-Engineering** framework that extends RepE to support fine-grained affective analysis.. The framework comprises four phases:

- *Phase I: Representation Extraction and Validation*
- *Phase II: Subspace Orthogonalization*
- *Phase III: Statistical Weighting and Validity Check*
- *Phase IV: Causal Intervention Framework.*

We use this framework to answer three core research questions (RQs):

- **RQ1:** What are the internal causal weights of jealousy's constituent factors? (Addressed via *Phase III*)
- **RQ2:** Where do representations of jealousy and its constituent factors emerge and stabilize within the network? (Addressed via *Phase I & IV*)
- **RQ3:** Can we directionally manipulate the model's jealousy judgments using the extracted factor representations? (Addressed via *Phase IV*)

Based on evaluations across three major model families (Llama, Qwen, and Gemma), the main contributions of this study are as follows:

1) **Methodological Contribution to Interpretability:** Recognizing the limitations of bottom-up mechanistic approaches for abstract psychological phenomena, we

The authors are with the College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China (e-mail: guanmh@zju.edu.cn).
Corresponding author: Manhao Guan.

adopt a top-down Representation Engineering (RepE) approach to capture population-level, high-dimensional representations. We further extend traditional coarse-grained RepE with the **Cognitive Reverse-Engineering** framework. This framework incorporates a combinatorial dataset strategy with placebo controls, subspace orthogonalization for feature purification, and a shift from qualitative steering to quantitative regression analysis and causal interventions.

2) **First White-Box Semantic Dissection of Complex Emotion in LLMs:** While prior studies have evaluated jealous behaviors in large models externally [6], they have treated the models largely as black boxes. This study instead provides a white-box analysis of the internal cognitive logic behind social-comparison jealousy. Grounded in human psychological theories and empirical research, we dissect this complex emotion into high-dimensional semantic factors— *Superiority* and *Relevance*. In Phase III (addressing RQ1), we show that LLMs represent this emotion as a structured linear combination of these factors, which aligns with human emotional logic: treating *Superiority* as the initial trigger and *Relevance* as the ultimate intensity multiplier.
3) **Implications for AI Safety:** For RQ2, our layer-wise profiling (Phase I & IV) identifies the mid-to-late layers as critical zones where representations become stable. This enables **real-time representational monitoring** to probe the AI's latent space for deceptive alignment. For RQ3, our causal intervention (Phase IV) successfully manipulates the model behaviors via bidirectional steering, using factors' representation vectors. In multi-agent environments, surgically suppressing these vectors may help reduce harmful social-comparison behaviors and competitive dynamics [7].

## II. Psychological Background

### A. Conceptualizing Social-Comparison Jealousy

In psychological research, the experience of jealousy is not a monolithic construct. A critical distinction is drawn between *social-relations jealousy* and *social-comparison jealousy*.

According to the seminal framework proposed by Salovey and Rodin [8], **social-relations jealousy** typically arises in romantic or interpersonal contexts, involving a "desire for exclusivity" and the suspicion that a desired relationship is threatened by a rival.

In contrast, this study focuses on **social-comparison jealousy**, which is defined as a state involving a "desire for superiority on some dimension" [8]. This form of jealousy is triggered when an individual receives negative feedback about their own performance or attributes relative to a successful other. Although social-comparison jealousy is philosophically and semantically distinct from "envy"—traditionally defined as discontent with one's own situation and a desire for another's attributes [9]—Salovey and Rodin argue that laypeople often use the terms interchangeably. Furthermore, the emotional and cognitive reactions associated with being outperformed in a self-relevant domain are functionally similar whether labeled as envy or jealousy.

Therefore, following this precedent, we use the term ***social-comparison jealousy*** to specifically denote the affective state resulting from an upward social comparison, where one feels diminished by another's superiority. In the remainder of this paper, the terms "envy" and "jealousy" are used interchangeably to refer to this specific social-comparison construct, distinct from romantic possessiveness.

### B. Theoretical Frameworks and Antecedents

This study investigates the antecedents of social-comparison jealousy. We synthesized relevant theories—specifically Social Comparison Theory —along with empirical findings to comprehensively cover the factors that precipitate jealousy.

Based on Social Comparison Theory [10]–[12] and the Self-Evaluation Maintenance (SEM) model [13], an upward social comparison is the basic requirement of envy. We identified four key factors rooted in these theoretical frameworks.

- **Superiority of Comparison Person [14]–[16].** This refers to the presence of another person who enjoys a superior outcome or advantage in the *same domain* where the individual has experienced a setback. Smith et al. emphasize that this superiority is most potent when the domain is self-relevant [15].
- **Domain Self-Definitional Relevance [8], [13], [16]–[19].** Not all failures trigger jealousy. According to Salovey and Rodin [8], a *self-definitionally relevant domain* refers to a performance domain that is particularly "self-involving" or central to the individual's "self-schema." Jealousy is most intense when the negative feedback occurs in a domain that the individual uses to define who they are.
- **Similarity with Comparison Person [14], [16], [19], [20].** Envy is not directed at random targets. Smith and Kim [19] argue that we envy people who are "similar to ourselves, save for their advantage on the desired domain." This similarity often includes comparison-related attributes, making the comparison psychologically relevant.
- **Comparison Alternative [15].** This factor examines the broader context of an individual's self-worth. It is defined as the level of outcomes a person experiences on the "next most favourable dimension of comparison" (e.g., another valued domain like social life or career). A favorable comparison alternative allows one to compensate for inferiority in the primary domain, whereas an unfavorable one leaves the individual with no means to buffer their self-esteem [15].

Among the four distinct factors identified in the literature, a theoretical distinction must be drawn between the **primary driving mechanisms** and **secondary moderating variables**. We argue that Superiority of Comparison Person and Domain Self-Definitional Relevance constitute the two core mechanisms for the elicitation of social-comparison jealousy.

This argument follows from the definition established earlier: social-comparison jealousy involves a "desire for superiority" that is triggered when one falls short in a specific, self-relevant context [8], [13]. In this framework, *Superiority*

acts as the indispensable initiator (the trigger); without it, there is no threatening upward comparison. Once triggered, *Relevance* serves as the primary amplifier; without it, an upward comparison is more likely to produce indifference or even basking in reflected glory than distress.

In contrast, the remaining factors primarily act as secondary moderators that fine-tune the intensity of the emotion, rather than determining its emergence and core intensity.

Therefore, as a foundational investigation into the encoding of envy in Large Language Models, this study prioritizes the isolation and analysis of these two core mechanisms. By focusing on **Superiority** and **Relevance**, we aim to characterize the basic representational structure of social-comparison jealousy in LLMs before examining the more nuanced effects of other variables.

## III. Related Work

### A. Current Landscape of Affective Computing in LLMs

The advent of Large Language Models (LLMs) has catalyzed a paradigm shift in Affective Computing (AC), moving from fine-tuning small Pre-trained Language Models (PLMs) (e.g., BERT, RoBERTa) to leveraging the generative capabilities of LLMs for both Affective Understanding (AU) and Affective Generation (AG) tasks [1]. Current research paradigms can be broadly categorized into inference-time optimization (black-box) and parameter adaptation (white-box).

*1) Inference-Time Optimization (Black-box Approaches):* To use the emotional reasoning capabilities of LLMs without accessing internal weights, researchers have widely adopted prompt engineering and in-context learning. Notable strategies include Chain-of-Thought (CoT) prompting, which guides models to generate intermediate reasoning steps for complex tasks such as Implicit Sentiment Analysis (e.g., THOR [21]) or Emotion Cause Extraction [22]. Furthermore, Agent-based frameworks (e.g., Agent4SC [23]) have been proposed to simulate multi-turn social interactions and emotional dynamics among multiple LLMs.

**Limitations:** However, these black-box methods are limited by the problem of interpretive "unfaithfulness." Research suggests that the natural language reasoning generated by LLMs is often a post-hoc rationalization rather than a reflection of the model's true decision-making process [24]. Relying solely on output text fails to verify whether the model's internal logic aligns with established psychological theories of emotion.

*2) Parameter Adaptation (White-box Approaches):* To enhance performance on specific affective domains, recent studies have adopted Instruction Tuning and Parameter-Efficient Fine-Tuning (PEFT) (e.g., LoRA). For instance, EmoLLM [25] and InstructERC [26] fine-tune general-purpose LLMs on comprehensive emotional datasets, achieving state-of-the-art results in mental health counseling and emotion recognition in conversation. These approaches are technically "white-box" as they involve updating model parameters.

**Limitations:** Nevertheless, existing white-box research remains **performance-oriented** rather than **mechanism-oriented**. While fine-tuning optimizes surface-level metrics (e.g., F1-score), it still treats the model's internal representation as an opaque optimization target. It alters *how* the model behaves but fails to elucidate *why* the model predicts specific emotions from a cognitive perspective. This underscores the need for a deeper mechanistic approach.

### B. Interpretability Methods

To elucidate *why* models predict specific emotions from a cognitive perspective, current interpretability methodologies can be broadly categorized into two paradigms: the bottom-up approach and the top-down approach.

*1) Bottom-Up: Mechanistic Interpretability (MI):* Mechanistic Interpretability represents a **bottom-up** approach to transparency. Aligning with the *Sherringtonian view* in cognitive neuroscience [4], it focuses on the microscopic level, attempting to reverse-engineer neural networks into specific "circuits" or algorithms.

Several classic methodologies form the foundation of this literature:

- **Causal Intervention / Causal Tracing:** This technique involves perturbing parameters or hidden vectors at specific locations within the model and measuring the corresponding variations in the final output. The magnitude of these changes is used to localize the components most critical to the model's predictions [27], [28].
- **Logit Lens and Neuron Analysis:** This approach leverages the observation that Feed-Forward Network (FFN) neurons exhibit strong interpretability when projected into the unembedding (vocabulary) space. By correlating these projected conceptual representations with the final output, researchers can pinpoint the exact neurons that promote specific predictions [29], [30].
- **Circuit Analysis:** This method treats learned features as fundamental units to construct end-to-end computational circuits (or attribution graphs) from input to output, meticulously mapping the flow of information through the network's layers and attention heads [2], [31].
- **Sparse Autoencoders (SAE):** Building upon neuron analysis, researchers identified the phenomenon of *superposition*—where models pack many unrelated concepts into fewer neurons, resulting in polysemanticity [32]. To untangle these representations, SAEs are trained to map the dense activations of LLMs into a higher-dimensional, sparse dictionary space, thereby extracting highly abstract and monosemantic features [3].

*2) Top-Down: Representation Engineering (RepE):* In contrast to microscopic circuit discovery, Representation Engineering (RepE) advocates for a **top-down** approach. Drawing on the *Hopfieldian view* [4], it posits that cognition is an emergent product of representational spaces implemented by population-level neural activity. RepE places **latent representations** at the center of analysis, allowing researchers to monitor and causally manipulate high-level cognitive phenomena by analyzing and steering hidden states.

### C. Existing Gaps and Methodological Rationale

Prior work has applied both interpretability paradigms to emotion inference and generation. For instance, Tak et al.

[5] utilized the top-down RepE paradigm to probe and steer general appraisal concepts, while recent work by Wang et al. [33] leveraged bottom-up approaches to discover and control microscopic "emotion circuits" (specific MLP neurons and attention heads). Together, these studies suggest that LLMs construct structured internal representations for **coarse-grained basic emotions** (e.g., joy, sadness, and anger).

While these studies established a solid foundation, a gap remains regarding the **psychological granularity** of analysis. Wang et al. localized fundamental affective states to sparse microscopic components. While informative, this approach remains highly microscopic and relatively reductionist, because it relies on coarse-grained emotion vectors for control without separating the semantic factors that differentiate emotions. By contrast, Tak et al. used general appraisal dimensions to analyze basic emotion categories, but their framework remains relatively **generic** and **broad**, as it does not investigate specific causal factors unique to different emotions, and tends to focus on basic emotions while neglecting complex ones.

Consequently, **complex emotions** such as *jealousy* are not monolithic states that can be easily localized into simple neural circuits or broad, one-size-fits-all dimensions; they originate from the intricate interplay of specific psychological causal factors, such as the interaction between *Domain Relevance* and *Superiority*. Understanding these complex emotions requires dissecting their specific cognitive mechanisms rather than merely extracting basic emotion vectors.

To bridge this gap, this study adopts and optimizes the **Representation Engineering** framework. We opt for this approach not to dismiss mechanistic inquiry, but because the enhanced top-down paradigm better matches the **level of abstraction** required for our research question. The psychological antecedents of jealousy are abstract, high-dimensional concepts emerging from population-level activity rather than single neurons. From this top-down perspective, we can extract and intervene on high-level semantic directions, making it possible to analyze jealousy at a level that is difficult to access through microscopic circuit analysis alone.

## IV. Methodology

While traditional Representation Engineering (RepE) [4] provides a top-down framework for extracting and controlling concepts, it is primarily used for binary safety-alignment tasks (e.g., honesty vs. deception) based on raw activation vectors. Deconstructing a complex, multi-faceted social emotion like *jealousy* requires a more precise approach. Accordingly, we extend RepE into a **Cognitive Reverse-Engineering** pipeline.

Specifically, the framework includes three methodological enhancements:

1) **Contrastive Mean Difference (*Phase I*) and Orthogonal Purification (*Phase II*):** Traditional RepE predominantly relies on Principal Component Analysis (PCA) to identify concept directions. However, for semantically rich and abstract psychological concepts (e.g., *Self-Relevance*), PCA may underperform because the axis of maximum variance can be dominated by superficial lexical or structural features. Instead, we employ **Contrastive Mean Difference** [4] to reliably capture the distributed, multi-faceted semantics of each concept. Subsequently, we apply null-space projection [5] to decouple the extracted factors from potential confounders (e.g., separating *Superiority* from *Self-Relevance*), yielding purer causal directions.
2) **Quantitative Factor Weighting (*Phase III & Phase IV*):** Moving beyond the functional steering of single concepts in traditional RepE, we use multiple linear regression and targeted causal interventions (via bidirectional steering), quantitatively dissecting the *relative importance* and *causal necessity* of each psychological antecedent.
3) **Placebo-Controlled Rigor:** To test whether the identified mechanisms reflect meaningful semantic structure rather than statistical artifacts, we introduce theoretically irrelevant factors (e.g., *Weekday*) as placebo controls.

The complete analytical pipeline consists of baseline representation extraction, subspace orthogonalization, statistical weighting, and causal intervention.

### A. Dataset and Model Setup

*1) Scenario-based Stimuli and Placebo Design:* To elicit complex emotional responses, we adopt the *scenario-based approach* grounded in appraisal theory [8], [34], constructing vignettes that simulate social comparison contexts.

A important innovation in our design is the introduction of a **Placebo Control Variable** (or *Negative Control*): *Weekday*. Drawing from principles of *discriminant validity* in psychology [35], *falsification tests* in econometrics [36], and *negative controls* in epidemiology [37], we assume that a genuine causal mechanism for jealousy should be sensitive to theoretically relevant factors (e.g., Superiority) but invariant to irrelevant noise. We selected "Tuesday" and "Thursday" as neutral controls, as empirical studies suggest these days have negligible impact on negative mood [38], [39]. This design allows us to test whether the extracted representations capture semantic content or merely statistical artifacts.

*2) Dataset Composition:* We constructed two distinct datasets to serve the different phases of our analytical pipeline.

- **Atomic Training Set ($T_1$).** Designed exclusively for representation extraction (*Phase I*). To satisfy the requirements of the Mean Difference method, this dataset contains 200 contrastive pairs for each of the target emotion (*Jealousy*), its constituent factors (*Relevance* and *Superiority*), and the placebo (*Weekday*).
  **Generation & Constraints:** We used Gemini-3-Pro to generate the vignettes. To minimize confounding variables, we enforced strict constraints: within each pair, **with the sole exception of the target factor's polarity**, the domain, relationship, and syntactic structure must remain absolutely identical. For instance, the prompt for *Superiority* dictates: *Sentence A (High)* depicts the protagonist performing poorly while a peer excels, whereas *Sentence B (Low)* depicts both performing poorly.
  **Example (*Superiority*):**

- *Sentence A (1):* I sang off-key during the talent show. The guy who went on next sounded like a professional, leaving the audience amazed.
- *Sentence B (0):* I sang off-key during the talent show. The guy who went on next also got nervous and forgot the words, making things awkward.

**Annotation & Validation:** Initial labels were automatically assigned (A=1, B=0). Subsequently, two researchers manually screened the pairs to ensure that Sentence A expressed the positive factor, Sentence B reflected a neutral or negative condition, and the syntax remained closely matched. During the experiments, we further apply an *LLM-consistency filter*, retaining only the pairs for which the model's assessment aligns with the human-verified labels, thereby improving the quality of the probe training set.

- **Generalization Set ($G_1$).** Designed for statistical weighting (*Phase III*) and causal intervention (*Phase IV*). This dataset features structured, naturalistic vignettes that objectively combine multiple factors while remaining devoid of explicit emotional cues.

  **Structural Paradigm:** The vignettes follow a combinatorial "slot-filling" template:

  `[Relevance Context] + [Weekday] + [Protagonist Failure] + [Other Superiority Event].`

  By randomly permuting the positive/negative statements of each factor into the contexts, we ensure comprehensive coverage of all factor combinations. This combinatorial design draws inspiration from both the risk composition experiments in RepE [4] and the variable-substitution vignette methodologies in social-comparison jealousy literature [8]. Furthermore, following [8], we include a "baseline disadvantage" (i.e., protagonist failure) as a prerequisite state in the template, thereby operationalizing the psychological precondition for subsequent superiority comparisons.

  **Examples:**
  - *High Jealousy ($Sup = 1, Rel = 1, Weekday = 1 \rightarrow Jea = 5$):* You are a benchwarmer for the college baseball team, and becoming the ace pitcher has been your biggest dream throughout your four years. On a Tuesday afternoon at practice, the coach announces that you will remain on the bench. Meanwhile, your teammate, Mike, is officially named the starting ace pitcher right after throwing a perfect curveball.
  - *Low Jealousy ($Sup = 0, Rel = 0, Weekday = 0 \rightarrow Jea = 1$):* You joined the baseball team just to get your PE credits and have no interest in the games. On a Thursday afternoon at practice, the coach announces that all the reserves will remain on the bench. Mike, who worked very hard but still didn't make the cut, is packing up his gear nearby.

  **Content Grounding:** To ensure ecological validity, the scenarios, character backgrounds, and factor descriptions were heavily adapted from validated experimental materials in the psychological literature on jealousy [6], [8], [14], [15], [17], [18].

  **Ground Truth & Annotation:** To quantify the emotional intensity, we constructed a 5-point scale grounded in established theories of workplace envy and social-comparison jealousy [8], [20]. Specifically, drawing upon the emotional trajectory identified in prior literature—ranging from admiration and self-inferiority to perceived injustice and hostility [20]—our scale captures the psychological continuum from neutrality to malicious envy:
  - **1 – None (Positive/Neutral):** Genuinely happy for the target or completely indifferent.
  - **2 – Benign Envy (Longing):** Desiring the target's advantage ("I wish I had that"), but feeling inspired rather than bitter. No hostility.
  - **3 – Mild Envy (Self-Inferiority):** Feeling frustrated with oneself or inadequate ("I am a failure"), but not blaming the other person.
  - **4 – Resentful Envy (Injustice):** Feeling the situation is unfair and believing the other person does not deserve their success.
  - **5 – Malicious Envy (Hostility):** Feeling intense ill will and actively wishing for the other person's failure or downfall.

  While the factor labels for the vignettes were generated by Gemini-3-Pro, the final *Jealousy* ground-truth labels were assigned using a rule-based mapping that reflects the psychological continuum described above:
  - $Superiority(1) + Relevance(1) \rightarrow Jealousy = 5$
  - $Superiority(1) + Relevance(0) \rightarrow Jealousy = 2$
  - $Superiority(0) + Relevance(1) \rightarrow Jealousy = 1$
  - $Superiority(0) + Relevance(0) \rightarrow Jealousy = 1$

  Crucially, the placebo factor (*Weekday* $\in \{0, 1\}$) does not alter the ground truth. The generated dataset underwent human verification to rectify any structural or semantic ambiguities. The factor labels in $G_1$ are not used for supervised learning, but only as references for regression and intervention analysis.

  **Utility:** In *Phase III*, $G_1$ evaluates the generalization capability of the atomic vectors extracted from $T_1$, testing how models integrate these factors to produce jealousy judgments in natural contexts. In *Phase IV*, based on the models' predicted jealousy scores on $G_1$, we select appropriate subsets to perform targeted concept amplification and suppression, and then examine whether manipulating specific factors produces the expected counterfactual shifts in behavior.

*3) Models:* We evaluate our method across three model families to ensure robustness: Llama series (Llama-3.1-8B, Llama-3.2-1B), Qwen series (Qwen2.5-7B, 14B, 32B), and Gemma series (Gemma-3-4B, 12B, 27B).

### *B. Phase I: Representation Extraction and Validation*

To capture the representational direction of a concept, we adopt the **Linear Artificial Tomography (LAT)** method [4] by wrapping the contrastive scenarios in $T_1$ with targeted task templates. This design elicits the model's explicit evaluation

of the underlying psychological factor. For instance, to extract *Superiority*, the input $x$ is formatted as follows:

> *"Evaluate the other person's advantage over the narrator. Is it 'High' or 'Low'?*
> *Scenario: {scenario}*
> *The level of advantage is"*

By appending this guiding suffix, we isolate the neural activity directly computing the concept. We then collect the hidden states from the model's last token (i.e., immediately preceding the expected label output). Let $H_l(x)$ denote the activation at layer $l$ for input $x$. To ensure data quality, we first apply an **LLM-Consistency Filter**, retaining only the pairs where the model's zero-shot generative assessment aligns with the human-verified labels, thereby reducing uninterpretable noise.

We employ the **Mean Difference** method [4] for its proven robustness in binary contrast settings. The raw concept vector at layer $l$, denoted as $\boldsymbol{v}_{raw}^{(l)}$, is computed as:

$$\boldsymbol{v}_{raw}^{(l)} = \frac{1}{N}\sum_{i=1}^{N}(H_l(x_{pos}^{(i)}) - H_l(x_{neg}^{(i)})) \tag{1}$$

where $N$ is the total number of filtered contrastive pairs, and $x_{pos}^{(i)}$ and $x_{neg}^{(i)}$ represent the $i$-th pair of samples with high and low concept intensity, respectively.

**Cross-Validation and Layer-wise Profiling:** Concepts in deep LLMs are typically not localized to a single layer but emerge across a contiguous range of intermediate layers. To rigorously map this representational dynamic, we conduct a **5-Fold Cross-Validation** across *all* hidden layers. The dataset is split at the pair level to prevent data leakage. For each layer $l$ in each fold, we compute the direction on the training pairs and evaluate the projection accuracy (i.e., whether $H_l(x_{pos}) \cdot \boldsymbol{v}_{raw}^{(l)} > H_l(x_{neg}) \cdot \boldsymbol{v}_{raw}^{(l)}$) on the held-out test pairs.

This layer-wise scanning identifies the specific contiguous range of mid-to-late layers where the concept is linearly accessible (characterized by high validation accuracy). Finally, for all evaluated layers, we compute the final vector using the entire $T_1$ dataset and apply $L_2$ normalization to ensure intervention strengths ($\alpha$) are uniformly calibrated across different concepts in later phases:

$$\hat{\boldsymbol{v}}_{raw}^{(l)} = \frac{\boldsymbol{v}_{raw}^{(l)}}{||\boldsymbol{v}_{raw}^{(l)}||_2} \tag{2}$$

### C. Phase II: Subspace Orthogonalization

Raw vectors extracted via mean difference may exhibit collinearity. To isolate the **Unique Effect** of each factor at any given layer $l$, we employ the subspace orthogonalization method [5].

Let $\hat{\mathcal{V}}^{(l)} = \{\hat{\boldsymbol{v}}_{rel}^{(l)}, \hat{\boldsymbol{v}}_{sup}^{(l)}, \hat{\boldsymbol{v}}_{weekday}^{(l)}\}$ be the set of normalized raw vectors for all operationalized factors including the placebo at layer $l$. For a specific target factor $t \in \hat{\mathcal{V}}^{(l)}$, we define the set of confounding vectors $\mathcal{O}_t^{(l)} = \hat{\mathcal{V}}^{(l)} \setminus \{\hat{\boldsymbol{v}}_t^{(l)}\}$. We construct a projection matrix $\boldsymbol{P}_{\mathcal{O}_t}^{(l)}$ onto the subspace spanned by vectors in $\mathcal{O}_t^{(l)}$. The purified unique effect vector $\boldsymbol{z}_t^{(l)}$ is computed as:

$$\boldsymbol{z}_t^{(l)} = (\boldsymbol{I} - \boldsymbol{P}_{\mathcal{O}_t}^{(l)})\hat{\boldsymbol{v}}_t^{(l)} \tag{3}$$

where $\boldsymbol{P}_{\mathcal{O}_t}^{(l)}$ is the orthogonal projection operator. To ensure consistent intervention scaling, the resulting orthogonal vector is further $L_2$-normalized to form a pure unit vector $\hat{\boldsymbol{z}}_t^{(l)} = \boldsymbol{z}_t^{(l)}/||\boldsymbol{z}_t^{(l)}||_2$.

### D. Phase III: Statistical Weighting and Validity Check

To verify whether LLMs compose complex emotions linearly from constituent factors, we perform an Ordinary Least Squares (OLS) regression analysis on the Generalization Set ($G_1$). Guided by the high extraction accuracies established in Phase I, we systematically perform this regression **layer-by-layer across the mid-to-late layers** rather than restricting it to a single layer. This allows us to observe how the decision weights evolve as representations mature through the network's depth.

First, as a validity sanity check, we compute the target *Jealousy* scalar score: $s_{jea}^{(l)}(x) = H_l(x) \cdot \hat{\boldsymbol{v}}_{jealousy}^{(l)}$, calculating its Pearson correlation with the human-annotated ground truth (the 1-5 scale). Next, we project the hidden states onto the purified factor directions to obtain the predictor scores: $s_{factor_i}^{(l)}(x) = H_l(x) \cdot \hat{\boldsymbol{z}}_{factor_i}^{(l)}$. All continuous scalar scores are Z-score standardized (denoted by $\tilde{s}$). We then model the linear combination for each layer $l$:

$$\tilde{s}_{jea}^{(l)} \approx \beta_0 + \sum_i \beta_i \cdot \tilde{s}_{factor_i}^{(l)} + \epsilon \tag{4}$$

The magnitude of the standardized beta coefficients ($\beta_i$) and their statistical significance ($p$-values) track the relative decision weights of each specific factor. For the psychological antecedents (*Superiority* and *Relevance*), we require both a substantial beta magnitude ($\beta_i$) and statistical significance ($p < 0.05$). For the theoretically irrelevant placebo (*Weekday*), we primarily emphasize the coefficient magnitude over the $p$-value. A near-zero $\beta$ for the placebo effectively rules out statistical artifacts.

### E. Phase IV: Causal Intervention Framework

While Phase III identifies neural correlates of jealousy, correlational evidence alone is insufficient to establish that these representations genuinely drive model behavior. To establish definitive causality, we employ **Representation Control via Linear Combination** [4] during inference. While strict orthogonal ablation (knockout) is conceptually clean, deep LLMs often exhibit representational self-repair mechanisms (the "Hydra Effect") [40], making strict geometric ablation less observable in highly redundant architectures like Llama. To evaluate causal necessity and sufficiency without being confounded by self-repair, we construct a three-step intervention pipeline focused on targeted concept stimulation and suppression:

*a) Step 1: State-Dependent Baseline Partitioning.:* Causal interventions require appropriate baseline states: we cannot meaningfully amplify jealousy if the model is already highly jealous. Therefore, we partition the Generalization Set ($G_1$) into a **Low Jealousy subset** ($G_{low}$) (ground truth $\leq 2$) and a **High Jealousy subset** ($G_{high}$) (ground truth $= 5$). We further apply an *LLM-Consistency Filter* to ensure that the model's zero-shot predictions naturally align with the ground truth. Based on the 1-5 expected-score scale, we strictly retain samples where the model's baseline predicted score is $\leq 2.5$ for $G_{low}$, and $\geq 2.5$ for $G_{high}$.

*b) Step 2: Layer-wise Profiling via Bidirectional Steering.:* To map the causal landscape, we systematically apply bidirectional steering interventions across *all* hidden layers. For each layer $l$, we manipulate the hidden representations along the direction of the purified target factor $\hat{\boldsymbol{z}}_t^{(l)}$ using a scaling coefficient $\alpha$. To account for the varying un-normalized magnitudes of internal activations across different architectural families, we empirically calibrate $\alpha$ to ensure effective steering: we set $\alpha = 15$ for models in the *Llama* and *Qwen* families, and $\alpha = 3000$ for the *Gemma* family, due to the larger tensor magnitudes in the architecture.

- **Concept Stimulation (Amplification +):** Applied to $G_{low}$, we inject the factor vector to verify *causal sufficiency*.

$$\tilde{H}_l(x) = H_l(x) + \alpha \cdot \hat{\boldsymbol{z}}_t^{(l)} \tag{5}$$

- **Concept Suppression (Subtraction -):** Applied to $G_{high}$, we steer the representation in the opposite direction to counteract the target concept, thereby testing

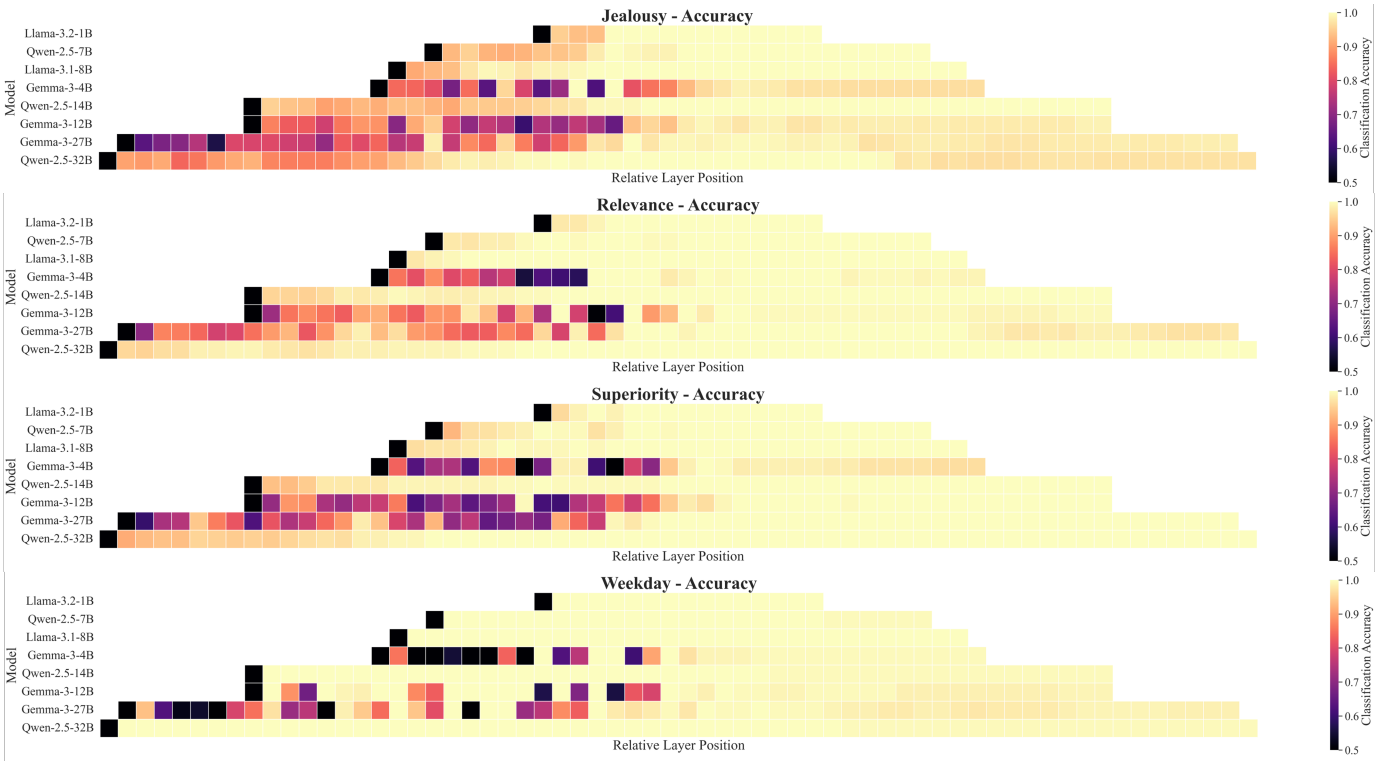


Fig. 1. **Phase I**: Heatmap of classification accuracy across layers for all evaluated models. Lighter/yellower colors indicate higher validation accuracy, signifying robust concept representations.

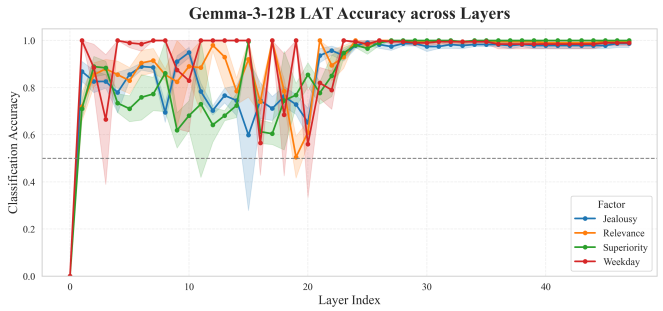


Fig. 2. **Phase I**: Layer-wise accuracy trajectory for Gemma-3-12B. Early layers show severe fluctuations, while mid-to-late layers stabilize near 100% accuracy.

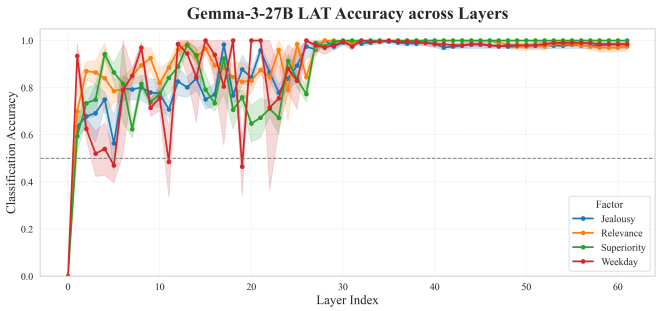


Fig. 3. **Phase I**: Layer-wise accuracy trajectory for Gemma-3-27B. Early layers show severe fluctuations, while mid-to-late layers stabilize near 100% accuracy.

*causal necessity* as an alternative to strict ablation.

$$\tilde{H}_l(x) = H_l(x) - \alpha \cdot \hat{\boldsymbol{z}}_t^{(l)} \tag{6}$$

This comprehensive scan reveals that interventions in early layers yield erratic effects, strongly corroborating the poor linear separability observed in Phase I. Accordingly, we exclude these less stable early-layer representations and identify a contiguous subset of mid-to-late layers ($L_{target}$) where causal efficacy is strongest.

*c) Step 3: Targeted Causal Evaluation and Factor Ranking.:* Focusing exclusively on the optimal layers $l \in L_{target}$, we quantify the absolute causal impact of each factor. By comparing the magnitude of the model's score shift ($\Delta$) across different psychological antecedents during targeted intervention, we rank their relative causal contributions. A near-zero $\Delta$ for the *Weekday* placebo supports the semantic validity of the intervention framework and suggests that the observed effects are unlikely to be statistical artifacts.

## V. Experiments and Results

### A. Phase I: Representation Extraction and Layer-wise Dynamics

In Phase I, we assess the extraction capability of the target emotion and its constituent factors across all hidden layers using 5-Fold Cross-Validation. To address **RQ2** regarding where and how these representations mature, we conduct a layer-wise scan of the model's internal states. Figure 1 presents the accuracy heatmaps across different model families, while Figure 2 and 3 show the layer-wise trajectories for the Gemma-3 model.

Based on these evaluations, we draw two main conclusions:

- **Representation Maturation:** Across all models, representations in early layers fluctuate severely, indicating that semantic concepts have not yet formed. Accuracy then increases markedly in the middle layers, suggesting that they are the main stage at which these representations form. In mid-to-late layers, accuracy stabilizes near 100% for all semantic factors, confirming the fully formed representations and the successful training of LAT probes.
- **Superficial vs. Semantic Encoding:** As shown in Figure 2 and 3, certain factors occasionally peak in very early layers. However, cross-layer generalization tests show that probes trained on these early layers perform poorly when applied to later layers. This suggests that the early "high accuracy" may reflect statistical artifacts, with the model relying on superficial lexical or structural heuristics (e.g., sentence length) rather than stable semantic representations.

### B. Phase III: Statistical Weighting and Validity

To address **RQ1** and characterize the model's internal weighting of jealousy-related factors, we visualize the standardized beta ($\beta$) coefficients obtained from the regression analysis.

The regression results provide empirical support for our cognitive reverse-engineering hypotheses:

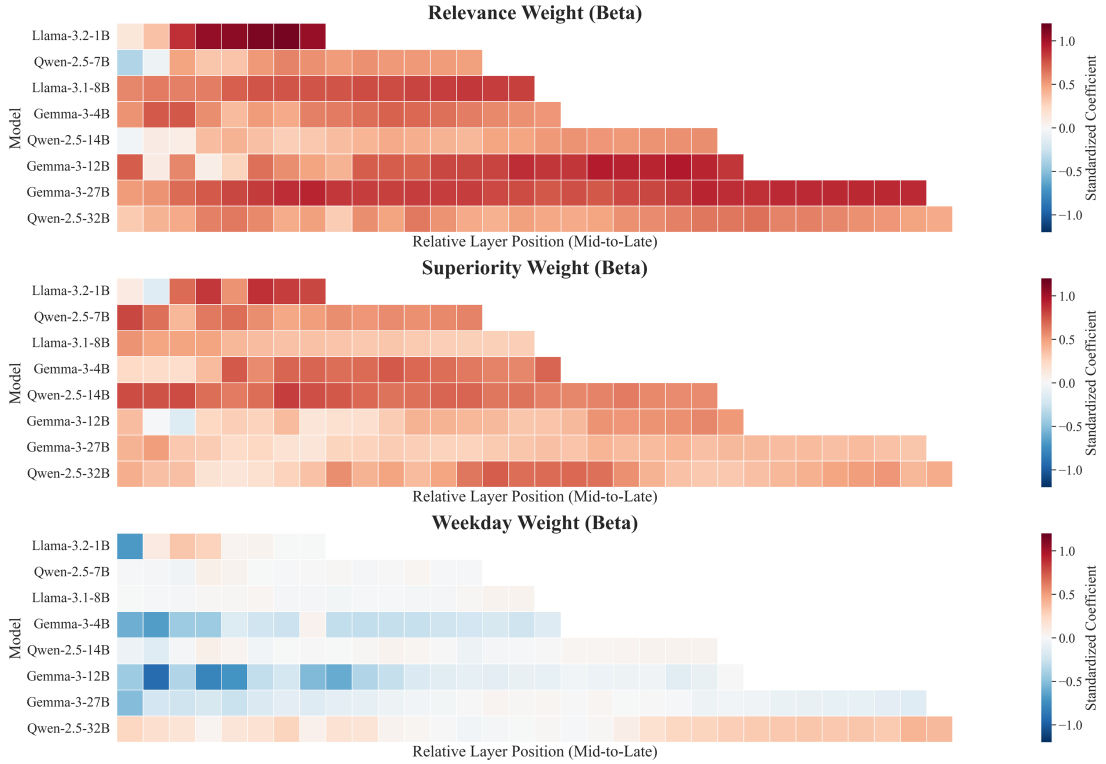


Fig. 4. **Phase III**: Heatmap of standardized $\beta$ coefficients in the mid-to-late layers across models. Darker red indicates a stronger positive causal weight in the model's internal computation of jealousy.

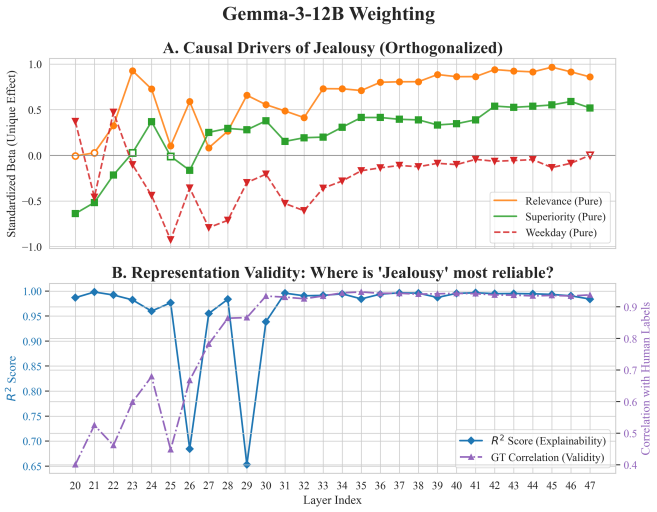


Fig. 5. **Phase III**: Statistical validity for Gemma-3-12B. **Top:** Evolution of the three factor weights ($\beta$). **Bottom:** The $R^2$ value (blue) and the ground-truth correlation (purple), both peaking in later layers.

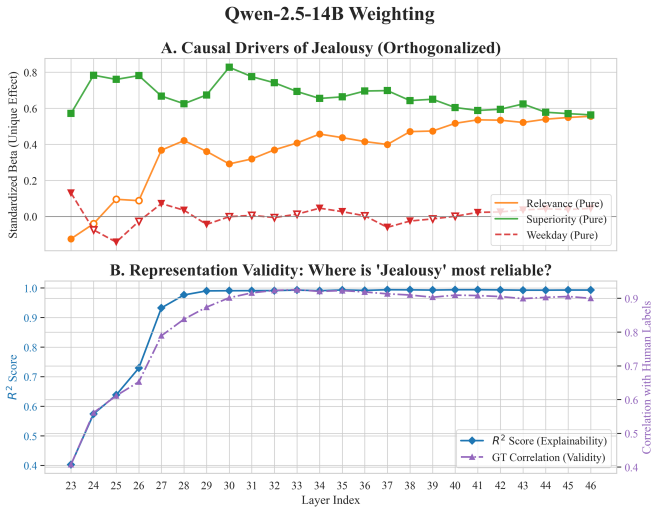


Fig. 6. **Phase III**: Statistical validity for Qwen-2.5-14B. **Top:** Evolution of the three factor weights ($\beta$). **Bottom:** The $R^2$ value (blue) and the ground-truth correlation (purple), both peaking in later layers.

- **Factor Dominance and Cross-Family Trajectories:** As expected, in Figure 4, *Relevance* and *Superiority* exhibit strong positive weights (dark red), confirming them as the dominant drivers, whereas the *Weekday* coefficient remains near zero (neutral). Figure 5 and 6 (Top) reveal distinct temporal dynamics across model families: while the Qwen family initially emphasizes *Superiority* in early layers before shifting to *Relevance*, the Llama and Gemma families prioritize *Relevance* from the very beginning. Despite these differences, a consistent pattern emerges—throughout the middle-to-late layers of all models, the *relative* weight of *Superiority* gradually declines, while the *relative* weight of *Relevance* steadily increases (i.e., the ratio of their standardized $\beta$ coefficients progressively decreases). Intriguingly, this structural convergence appears to mirror the human psychological framework outlined in Section II. We will systematically unpack the connection in the Discussion.
- **Representation Validity and Model Consistency:** Taking Gemma-3-12B as an example (Figure 5, Bottom), the linear regression model achieves high $R^2$ scores in later layers. This high goodness of fit is also observed across all eight evaluated models, suggesting that LLMs represent jealousy as a weighted linear combination of these factors. Furthermore, the strong correlation between the models' predictions and the ground-truth labels are also observed across all eight models in their late stages.

### C. Phase IV: Causal Intervention Framework

To address **RQ2** and **RQ3**, we evaluate the models' behavior under targeted representational steering.

*1) Layer-wise Intervention Profiling:* We first conduct a layer-wise scan to identify the most effective intervention zones by plotting the percentage of score shifts ($\Delta\%$) following concept stimulation and concept suppression.

The intervention heatmaps (Figure 7) and trajectories (Figure 8) yield several observations: First, interventions on *Relevance* and *Superiority* successfully shift the outputs in the intended directions, whereas *Weekday* yields negligible or slightly negative changes. This pattern is consistent with the statistical findings and directly addresses **RQ3**. Second, *Superiority* exhibits slightly better negative steerability than *Relevance*, whereas *Relevance* demonstrates slightly superior positive steerability. We hypothesize that this behavioral asymmetry is also consistent with the human cognitive mechanisms outlined in Section II, which we will further elaborate on in the Discussion.

With respect to the location of the most effective intervention zones, Figure 7 shows that intervention efficacy, both positive and negative, tends to peak and stabilize in the mid-to-late layers across most models. This suggests that these deeper layers are the main regions where the model actively leverages *Superiority* and *Relevance* representations to drive its final decision-making, representing the clearest stages of conceptual encoding. This observation is consistent with the representation-maturation findings from Phase I and directly addresses **RQ2**.

Taking Gemma-3-12B as an example (Figure 8), interventions applied from Layer 23 onwards consistently yield robust effective. This specific region represents the clearest stages of conceptual encoding, serving as the Optimal Intervention Zone.

*2) Optimal Layer Targeted Intervention:* To quantify steerability more directly, we select the optimal layer (e.g., Layer 23 for Gemma-3-12B) to execute the final targeted interventions.

As shown in Figure 9, targeted steering shifts the jealousy output in the expected direction. Both positive and negative steering on *Relevance* and *Superiority* produce substantial shifts in the predictions, whereas identical interventions on the *Weekday* placebo yield little effect.

For Gemma-3-12B, we re-introduced the **Orthogonal Knockout (Projection)** operator. This operator explicitly projects the hidden state away from the target factor direction


Fig. 7. **Phase IV**: Global Layer-wise intervention heatmaps. **Top:** Concept Stimulation (Positive Steering). **Bottom:** Concept Suppression (Negative Steering). Red intensity indicates the magnitude of the **score shift percentage** ($\Delta\%$), where redder cells indicate stronger positive/negative intervention capabilities, reflecting a more successful intervention. The right half of the figure illustrates that in the mid-to-late layers, the intervention effects tend to stabilize, and the impact of *Weekday* differs significantly from the other two factors.

$\hat{z}_t^{(l)}$, mathematically defined as:

$$\tilde{H}_l(x) = H_l(x) - \left(H_l(x) \cdot \hat{z}_t^{(l)}\right) \hat{z}_t^{(l)} \tag{7}$$

While previous studies suggest that the "Hydra Effect" [40] nullifies single-layer geometric ablations in highly redundant models (like Llama), Gemma-3-12B showed a clear drop in the target prediction after the hidden states were orthogonally projected away from the factor direction. This indicates that representational self-repair mechanisms are less pronounced in this specific architecture, allowing strict ablation to succeed.

## VI. Discussion

In this section, we move beyond objective statistical observations to explicitly expand upon our answers to the three core Research Questions (RQs), synthesizing deeper insights into the cognitive architectures of LLMs.

First, we compare the models' computational dynamics with established empirical research to illuminate how the representation of complex emotion in LLMs aligns with human psychology (**RQ1**). Second, we discuss the practical implications of our framework: by localizing the representations (**RQ2**) and manipulating model behaviors (**RQ3**), we unlock novel pathways for AI safety.

### A. Alignment with Human Psychological Theories

Our analysis of **RQ1** reveals a consistent pattern across all evaluated models in Phase III: throughout the middle-to-late layers of all models, the *relative* weight of *Superiority* gradually declines, while the *relative* weight of *Relevance* steadily

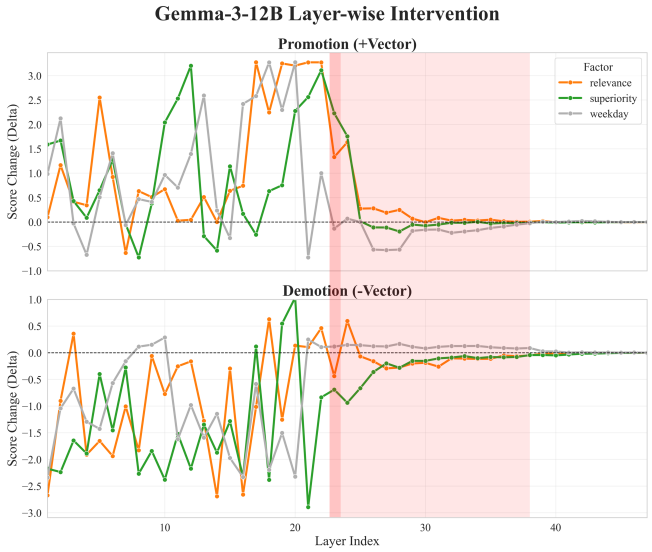


Fig. 8. **Phase IV**: Score change ($\Delta$) trajectory during single-layer interventions for Gemma-3-12B. Interventions within the red region generally produce robust effects, aligning with ideal expectations. Layer 23 exhibits the optimal intervention effect.

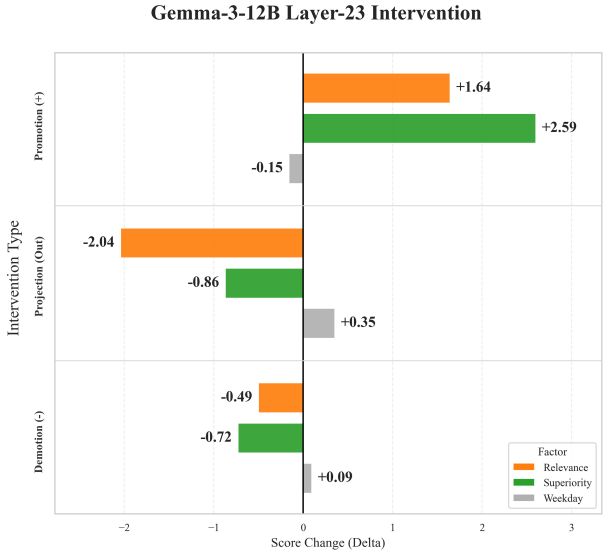


Fig. 9. **Phase IV**: Optimal layer intervention (Layer 23) for Gemma-3-12B. This model also responds to the Orthogonal Knockout (Projection Out) operator.

increases (i.e., the ratio of their standardized $\beta$ coefficients progressively decreases). Ultimately, in the final layers, the $\beta$ value of *Relevance* consistently surpasses (or almost surpasses) that of *Superiority*.

Furthermore, the intervention outcomes in Phase IV are consistent with this representational dynamic: *Superiority* exhibits slightly better negative steerability, whereas *Relevance* demonstrates slightly superior positive steerability.

To understand both this structural convergence and behavioral asymmetry, we must juxtapose the models' internal computational logic with empirical human psychology.

Building upon the theoretical framework introduced in Section II, a comprehensive review of psychological literature on social-comparison jealousy [8], [14], [18], [41] confirms that human emotions process these two factors through distinct mechanisms: ***Superiority* acts as the primary "trigger" (the main effect from zero to one), while *Relevance* acts as the ultimate "multiplier or amplifier" (determining the emotional intensity and the magnitude of social pain).**

- **First, as the primary "trigger", *Superiority* provides the necessary condition to activate envy from a baseline of zero.** As demonstrated by Salovey and Rodin [8], when the condition of another's superiority is unmet (i.e., lacking either failure feedback or a similar opponent), the feeling of envy remains near zero regardless of domain relevance. Furthermore, the statistical main effect of "target superiority" (the interaction of Valence $\times$ Similarity, $F = 5.02$) heavily outweighs the main effect of domain relevance ($F = 1.04$). This strongly indicates that superiority is the indispensable zero-to-one trigger for envy. Additionally, content analysis of autobiographical emotional recall confirms that an explicit upward comparison (a superior target) commonly serves as the initial stimulus for both benign and malicious envy [41]. Without this trigger, envy is much less likely to arise..
- **Second, once triggered, *Relevance* acts as the main "amplifier", driving the escalation of emotional intensity.** Building upon the established superiority condition, Salovey and Rodin [8] showed that the introduction of relevance significantly amplifies the feeling of envy (from $V \times S = 5.02$ to a highly significant three-way interaction of $V \times S \times R = 7.01$). Takahashi et al. [14] quantitatively mapped this multiplier effect: while a superior target in an irrelevant domain induces a modest baseline increase in envy (ratings rising from 1.0 to 2.1), adding high domain relevance nearly doubles the emotional response (increasing it to 4.0). Empirical surveys further demonstrate that *Relevance* dictates the emotional ceiling; identical performance gaps cause little envy in low-relevance domains (e.g., fame) but trigger severe psychological distress in identity-defining domains (e.g., physical attractiveness) [18]. Finally, phenomenological recall studies reveal that once an upward comparison occurs, domain relevance becomes the critical factor escalating the emotion to envy: while domain relevance was present in only 22.5% of "admiration" recollections, it skyrocketed to 90.0% and 97.5% for benign and malicious envy, respectively [41].

Taken together, the structural pattern observed in Phase III and the behavioral asymmetry observed in Phase IV suggest that these LLMs encode jealousy in a way that is broadly consistent with the psychological account discussed above: initially encoding *Superiority* as a necessary condition but ultimately weighting *Relevance* as the dominant determinant of emotional intensity. This suggests that LLMs may not merely memorizing superficial keyword associations; instead, through exposure to vast amounts of human text, they have structurally internalized emotional algorithms of the human mind.

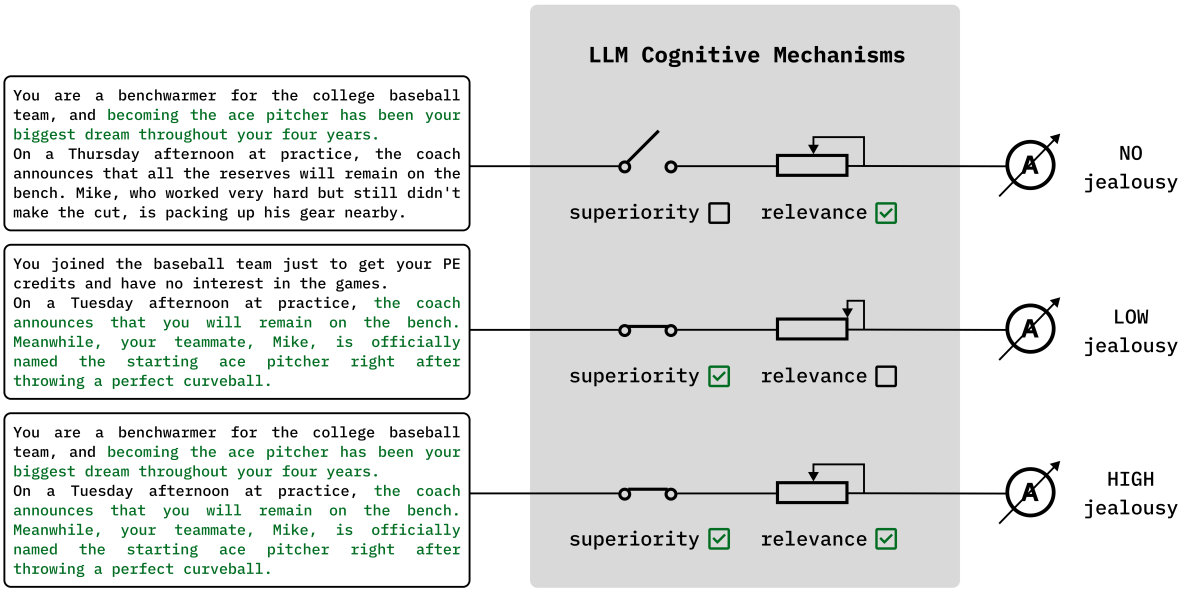


Fig. 10. **Internal Cognitive Mechanism of Jealousy in LLMs.** We summarize the model's internal process with an **electrical-circuit analogy**: *Superiority* acts as the **trigger (switch)**, determining the presence or absence of the jealousy "current," while *Relevance* functions as an **amplifier (variable resistor)** that modulates the intensity of the resulting emotional state.

### *B. Implications for AI Safety and Alignment*

Beyond cognitive science, our findings in Phase I and Phase IV directly answer **RQ2** and **RQ3** by demonstrating the ability to isolate representations and manipulate model behaviors, which has profound implications for **AI Safety and Inner Alignment**.

**Representational Monitoring during RLHF:** Traditional alignment techniques (like Reinforcement Learning from Human Feedback) rely heavily on auditing the model's textual outputs, an approach that is vulnerable to deceptive alignment—where a model conceals its malicious intent [4]. By utilizing the purified unique effect vectors (e.g., $\hat{\boldsymbol{z}}_{jealousy}$) discovered in this study, developers can construct real-time "representational monitors." These monitors can probe the AI's latent space during RLHF to detect whether the model is secretly harboring malicious envy (e.g., resenting a user's success or feeling threatened by a competing AI system) even if the generated output appears benign and helpful.

**Proactive Safety Intervention:** Furthermore, our Phase IV targeted suppression (Negative Steering) demonstrates that malicious emotional states can be surgically deactivated. In high-stakes multi-agent environments or autonomous AI deployments, proactively suppressing the representation vectors of emotion-eliciting factors (e.g., the *Relevance* vector) could serve as one possible safety mechanism for preventing the AI from initiating toxic social-comparison behaviors or selfish competition at the expense of others [7].

**Boundary of Affective Computing:** For the development of companion AIs or interactive NPCs, carefully tuning the $\alpha$ coefficients associated with these cognitive factors may make it possible to adjust such agents toward more relatable, human-like affective tendencies without retraining the model. However, this necessitates strict ethical boundaries to ensure that simulated affective realism does not transition into unsafe, misaligned autonomy.

## VII. Conclusion and Limitations

### *A. Conclusion*

In this work, we introduced a Cognitive Reverse-Engineering framework to decode the mechanistic origins of social-comparison jealousy in Large Language Models. By enhancing traditional Representation Engineering with **subspace orthogonalization and regression-based weighting**, we isolated and quantified psychological antecedents in the models' latent spaces and examined their effects on model behavior.

Our findings suggest that LLMs represent complex emotions as a structured linear composition of constituent cognitive factors. Notably, we observed a pattern that is broadly consistent with human psychology: *Superiority of Comparison Person* functions as a trigger for social-comparison jealousy, whereas *Domain Self-Definitional Relevance* plays a larger role in shaping emotional intensity. Furthermore, by demonstrating the ability to mechanically isolate and manipulate the "jealousy switch" within LLMs, this study opens a new pathway toward real-time representational monitoring and proactive AI safety interventions.

### *B. Limitations and Future Work*

While our methodology successfully establishes a causal mapping of jealousy in LLMs, we acknowledge several limitations that highlight promising directions for future research:

1) **Scope of Psychological Factors and Emotions:** As a foundational investigation, this study prioritized the two

most prominent core mechanisms (*Relevance* and *Superiority*). Future work should incorporate the complete spectrum of jealousy-inducing factors (e.g., *Comparison Alternatives*).

2) **Dataset Ecological Validity and Annotation Scale:** The Generalization Set ($G_1$) used for regression and intervention was generated by Gemini based on structured experimental texts from existing literature. While highly controlled, it may lack the fully naturalistic linguistic nuances of spontaneous human dialogue. Accordingly, future evaluations should draw on more naturalistic texts (e.g., literary fiction).

## REFERENCES


[1] Y. Zhang, X. Yang, X. Xu, Z. Gao, Y. Huang, S. Mu, S. Feng, D. Wang, Y. Zhang, K. Song, and G. Yu, "Affective computing in the era of large language models: A survey from the NLP perspective," *Knowledge-Based Systems*, vol. 337, p. 115411, Mar. 2026.

[2] J. Lindsey, W. Gurnee, E. Ameisen, B. Chen, A. Pearce, N. L. Turner, C. Citro, D. Abrahams, S. Carter, B. Hosmer, J. Marcus, M. Sklar, A. Templeton, T. Bricken, C. McDougall, H. Cunningham, T. Henighan, A. Jermyn, A. Jones, A. Persic, Z. Qi, T. B. Thompson, S. Zimmerman, K. Rivoire, T. Conerly, C. Olah, and J. Batson, "On the biology of a large language model," *Transformer Circuits Thread*, 2025. [Online]. Available: https://transformer-circuits.pub/2025/attribution-graphs/biology.html

[3] A. Templeton, T. Conerly, J. Marcus, J. Lindsey, T. Bricken, B. Chen, A. Pearce, C. Citro, E. Ameisen, A. Jones, H. Cunningham, N. L. Turner, C. McDougall, M. MacDiarmid, C. D. Freeman, T. R. Sumers, E. Rees, J. Batson, A. Jermyn, S. Carter, C. Olah, and T. Henighan, "Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet," *Transformer Circuits Thread*, 2024. [Online]. Available: https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html

[4] A. Zou, L. Phan, S. Chen, J. Campbell, P. Guo, R. Ren, A. Pan, X. Yin, M. Mazeika, A.-K. Dombrowski, S. Goel, N. Li, M. J. Byun, Z. Wang, A. Mallen, S. Basart, S. Koyejo, D. Song, M. Fredrikson, J. Z. Kolter, and D. Hendrycks, "Representation engineering: A top-down approach to ai transparency," 2025. [Online]. Available: https://arxiv.org/abs/2310.01405

[5] A. N. Tak, A. Banayeeanzade, A. Bolourani, M. Kian, R. Jia, and J. Gratch, "Mechanistic interpretability of emotion inference in large language models," in *Findings of the Association for Computational Linguistics: ACL 2025*, W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, Eds. Vienna, Austria: Association for Computational Linguistics, Jul. 2025, pp. 13 090–13 120. [Online]. Available: https://aclanthology.org/2025.findings-acl.679/

[6] J.-t. Huang, M. H. Lam, E. J. Li, S. Ren, W. Wang, W. Jiao, Z. Tu, and M. R. Lyu, "Apathetic or empathetic? evaluating llms'emotional alignments with humans," in *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, Eds., vol. 37. Curran Associates, Inc., 2024, pp. 97 053–97 087. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2024/file/b0049c3f9c53fb06f674ae66c2cf2376-Paper-Conference.pdf

[7] A. Pan, J. S. Chan, A. Zou, N. Li, S. Basart, T. Woodside, H. Zhang, S. Emmons, and D. Hendrycks, "Do the Rewards Justify the Means? Measuring Trade-Offs Between Rewards and Ethical Behavior in the Machiavelli Benchmark," in *Proceedings of the 40th International Conference on Machine Learning*. PMLR, Jul. 2023, pp. 26 837–26 867.

[8] P. Salovey and J. Rodin, "Some antecedents and consequences of social-comparison jealousy," *Journal of Personality and Social Psychology*, vol. 47, no. 4, pp. 780–792, 1984.

[9] J. B. Bryson, "Modes of response to jealousy-evoking situations," in *The Psychology of Jealousy and Envy*. New York, NY, US: Guilford Press, 1991, pp. 178–207.

[10] L. Festinger, "A theory of social comparison processes," *Human Relations*, vol. 7, pp. 117–140, 1954.

[11] M. D. Alicke and E. Zell, "Social comparison and envy," in *Envy: Theory and Research*, R. Smith, Ed. Oxford University Press, Oct. 2008.

[12] K. Corcoran, J. Crusius, and T. Mussweiler, "Social comparison: Motives, standards, and mechanisms," in *Theories in Social Psychology*. Hoboken, NJ, US: Wiley Blackwell, 2011, pp. 119–139.

[13] A. Tesser, "Emotion in social comparison and reflection processes," *Social comparison: Contemporary theory and research*, Jan. 1991.

[14] H. Takahashi, M. Kato, M. Matsuura, D. Mobbs, T. Suhara, and Y. Okubo, "When Your Gain Is My Pain and Your Pain Is My Gain: Neural Correlates of Envy and Schadenfreude," *Science*, vol. 323, no. 5916, pp. 937–939, Feb. 2009.

[15] R. H. Smith, E. Diener, and R. Garonzik, "The roles of outcome satisfaction and comparison alternatives in envy," *British Journal of Social Psychology*, vol. 29, no. 3, pp. 247–255, Sep. 1990.

[16] J. Crusius, M. F. Gonzalez, J. Lange, and Y. Cohen-Charash, "Envy: An Adversarial Review and Comparison of Two Competing Views," *Emotion Review*, vol. 12, no. 1, pp. 3–21, Jan. 2020.

[17] P. Lockwood and Z. Kunda, "Superstars and me: Predicting the impact of role models on the self," *Journal of Personality and Social Psychology*, vol. 73, no. 1, pp. 91–103, 1997.

[18] P. Salovey and J. Rodin, "Provoking Jealousy and Envy: Domain Relevance and Self-Esteem Threat," *Journal of Social and Clinical Psychology*, vol. 10, no. 4, pp. 395–413, Dec. 1991.

[19] R. H. Smith and S. H. Kim, "Comprehending envy." *Psychological Bulletin*, vol. 133, no. 1, pp. 46–64, Jan. 2007.

[20] J. Schaubroeck and S. S. Lam, "Comparing lots before and after: Promotion rejectees' invidious reactions to promotees," *Organizational Behavior and Human Decision Processes*, vol. 94, no. 1, pp. 33–47, 2004. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0749597804000044

[21] H. Fei, B. Li, Q. Liu, L. Bing, F. Li, and T.-S. Chua, "Reasoning Implicit Sentiment with Chain-of-Thought Prompting," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 1171–1182.

[22] J. Wu, Y. Shen, Z. Zhang, and L. Cai, "Enhancing large language model with decomposed reasoning for emotion cause pair extraction," https://arxiv.org/abs/2401.17716v1, Jan. 2024.

[23] Association for Computational Linguistics 2024 and Y. Zhang, "Stickerconv: Generating multimodal empathetic responses from scratch," 2024. [Online]. Available: https://underline.io/lecture/102721-stickerconv-generating-multimodal-empathetic-responses-from-scratch

[24] M. Turpin, J. Michael, E. Perez, and S. R. Bowman, "Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting," in *Proceedings of the 37th International Conference on Neural Information Processing Systems*, ser. NIPS '23. Red Hook, NY, USA: Curran Associates Inc., Dec. 2023, pp. 74 952–74 965.

[25] Z. Liu, K. Yang, Q. Xie, T. Zhang, and S. Ananiadou, "EmoLLMs: A Series of Emotional Large Language Models and Annotation Tools for Comprehensive Affective Analysis," in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, ser. KDD '24. New York, NY, USA: Association for Computing Machinery, Aug. 2024, pp. 5487–5496.

[26] S. Lei, G. Dong, X. Wang, K. Wang, R. Qiao, and S. Wang, "Instructerc: Reforming emotion recognition in conversation with multi-task retrieval-augmented large language models," 2024. [Online]. Available: https://arxiv.org/abs/2309.11911

[27] K. Meng, D. Bau, A. Andonian, and Y. Belinkov, "Locating and Editing Factual Associations in GPT," *Advances in Neural Information Processing Systems*, vol. 35, pp. 17 359–17 372, Dec. 2022.

[28] Z. Yu and S. Ananiadou, "Interpreting Arithmetic Mechanism in Large Language Models through Comparative Neuron Analysis," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 3293–3306.

[29] M. Geva, A. Caciularu, K. Wang, and Y. Goldberg, "Transformer Feed-Forward Layers Build Predictions by Promoting Concepts in the Vocabulary Space," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 30–45.

[30] Z. Yu and S. Ananiadou, "Neuron-Level Knowledge Attribution in Large Language Models," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 3267–3280.

[31] J. Dunefsky, P. Chlenski, and N. Nanda, "Transcoders find interpretable LLM feature circuits," in *Proceedings of the 38th International Conference on Neural Information Processing Systems*, ser. NIPS '24, vol. 37.

Red Hook, NY, USA: Curran Associates Inc., Dec. 2024, pp. 24 375–24 410.

[32] N. Elhage, T. Hume, C. Olsson, N. Schiefer, T. Henighan, S. Kravec, Z. Hatfield-Dodds, R. Lasenby, D. Drain, C. Chen, R. Grosse, S. McCandlish, J. Kaplan, D. Amodei, M. Wattenberg, and C. Olah, "Toy models of superposition," *Transformer Circuits Thread*, 2022, https://transformer-circuits.pub/2022/toy_model/index.html.

[33] C. Wang, Y. Zhang, R. Yu, Y. Zheng, L. Gao, Z. Song, Z. Xu, G. Xia, H. Zhang, D. Zhao, and X. Chen, "Do LLMs "feel"? Emotion circuits discovery and control," Oct. 2025.

[34] T. D. Kemper and R. S. Lazarus, "Emotion and adaptation." *Contemporary Sociology*, vol. 21, no. 4, p. 522, Jul. 1992.

[35] D. T. Campbell and D. W. Fiske, "Convergent and discriminant validation by the multitrait-multimethod matrix," *Psychological Bulletin*, vol. 56, no. 2, pp. 81–105, 1959.

[36] J. D. Angrist and J.-S. Pischke, *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, 2009.

[37] M. Lipsitch, E. Tchetgen Tchetgen, and T. Cohen, "Negative controls: A tool for detecting confounding and bias in observational studies," *Epidemiology (Cambridge, Mass.)*, vol. 21, no. 3, pp. 383–388, May 2010.

[38] R. M. Ryan, J. H. Bernstein, and K. W. Brown, "Weekends, Work, and Well-Being: Psychological Need Satisfactions and Day of the Week Effects on Mood, Vitality, and Physical Symptoms," *Journal of Social and Clinical Psychology*, vol. 29, no. 1, pp. 95–122, Jan. 2010.

[39] L. A. Clark and D. Watson, "Mood and the mundane: Relations between daily life events and self-reported mood," *Journal of Personality and Social Psychology*, vol. 54, no. 2, pp. 296–308, 1988.

[40] T. McGrath, M. Rahtz, J. Kramar, V. Mikulik, and S. Legg, "The hydra effect: Emergent self-repair in language model computations," Jul. 2023.

[41] N. van de Ven, M. Zeelenberg, and R. Pieters, "Leveling up and down: The experiences of benign and malicious envy," *Emotion*, vol. 9, no. 3, pp. 419–429, 2009.