Journal of High School Science

Emotion-associated representations in the feed forward gates of a language model

Upadhyay R[1], Balasubramanian S[2]

Submitted: June 24, 2026, Revised: version 1, July 27, 2026, version 2, August 6, 2026, version 3, August 24, 2026, version 4, October 6, 2026, version 5, October 13, 2026
Accepted: October 19, 2026

Abstract

Identifying emotion-associated neurons in large language models does not establish how those neurons contribute to emotional behavior. This study investigates that distinction through three levels of explanation: representation, reconfiguration, and causal control. Statistically corrected feed-forward activation signatures in Qwen2.5-1.5B-Instruct exhibited structured overlap corresponding to a psychological circumplex model, with substantial but incomplete reproducibility across Chinese texts and their English translations. Four experimental task families investigated whether these signatures changed with behavioral demands and predicted model outputs. The preregistered internal measure failed held-out validation, with only two of fourteen hypotheses supported; post hoc corrected measures revealed task-dependent reconfiguration and several associations with behavioral measures. However, increased internal separation did not consistently improve discrimination accuracy, evidence-specific changes depended on the internal measure, and static neuronal overlap did not predict generated-text confusability under the designated classifier, although a secondary lexicon analysis showed the predicted association. Suppression of emotional expression was not accompanied by a reliably detected change at the sampled pre-generation position. These findings demonstrate that emotion-associated activation structure and task responsiveness do not, by themselves, provide a general explanation of emotional behavior. The study establishes the need to distinguish representational association, task-dependent reconfiguration, and causal control, the last of which remains untested.

Keywords

Large Language Models, Mechanistic interpretability, Emotion representation, Artificial neural networks, Transformer neural networks, Feed-forward neural networks, Neural activation patterns, Representational reconfiguration, Causal inference, AI interpretability

[1]Rayan Upadhyay, The International School Bangalore (TISB), NAFL Valley Circle, Sarjapur - Marathahalli Rd, near Dommasandra, Whitefield, Thigala Chowdadenahalli, Karnataka 562125, India.
[2]Corresponding author: Sankar Balasubramanian, Indian Institute of Science (IISc), C V Raman Avenue, Bangalore, Karnataka 560012, India. sankarb@iisc.ac.in

https://doi.org/10.64336/001c.175619

1. Introduction

1.1 Background and motivation

Large language models (LLMs) can recognize and generate emotionally expressive text, yet the internal computations underlying these capabilities remain incompletely understood. In particular, it is unclear whether different emotions are associated with identifiable neural representations, how these representations relate to one another, and whether their organization changes with the task being performed. Behavioral performance alone cannot answer these questions.

Transformer feed-forward networks (FFNs), including their gating mechanisms, provide an accessible location for examining emotion-associated neural activity. Because emotions may share linguistic and semantic features, their representations need not occupy exclusive neuronal populations. Identifying patterns of activation and overlap may therefore reveal how emotional information is organized internally, but does not establish that the identified neurons determine the model's behavior.

This study distinguishes three levels of investigation: *representation*, the identification of emotion-associated neural activity; *reconfiguration*, changes in that activity under different task demands; and *causal control*, the mechanisms through which neural computations influence behavioral outcomes. It investigates representation and reconfiguration and tests whether their measured properties correspond to observable behavior, without assuming that statistical association establishes causal control.

1.2 Research questions

The study addresses three principal research questions.

RQ1 — Representation: Can statistically distinguishable neural activation signatures be identified for six primary emotions in the feed-forward gates of a language model, and do these signatures exhibit reproducible patterns of overlap that correspond to established relationships between emotions?

RQ2 — Reconfiguration: Do these emotion-associated representations change systematically when the model performs different emotional tasks, including recognition, discrimination, evidence-based disambiguation, targeted emotional expression, and suppression?

RQ3 — Behavioral significance: Do the measured organization and task-dependent changes of these representations predict the model's observable emotional behavior, including discrimination accuracy, behavioral confusability, and emotional expression?

Together, these questions distinguish the existence of emotion-associated representations from their task-dependent variation and their proposed behavioral significance. Establishing whether the identified representations are causally necessary or sufficient for particular behaviors remains outside the experimental scope of this study.

2. Related work

2.1 Emotion representation in Natural Language Processing

Computational approaches to emotion recognition in text have evolved from lexicon-based methods, such as the NRC Word-Emotion

Association Lexicon (1), to machine learning classifiers, convolutional neural networks (2), and transformer-based language models (3). These approaches commonly employ categorical frameworks, particularly Ekman's six basic emotions—anger, disgust, fear, happiness, sadness, and surprise (4). Alternative frameworks include Plutchik's wheel of emotions (5) and dimensional models based on valence, arousal, and dominance (6, 7). More recent research has introduced finer-grained emotion taxonomies and larger annotated datasets (8), while constructionist theories challenge the assumption that emotions constitute fixed, biologically distinct categories, instead emphasizing their dependence on context and underlying affective processes (9).

Despite these developments, much of the research has focused on behavioral performance, particularly the accuracy with which models classify or generate emotionally expressive text. Comparatively less attention has been devoted to identifying the internal neural representations associated with different emotions, their potential overlap, and their responses to changing task demands.

2.2 Mechanistic interpretability and knowledge neurons

Mechanistic interpretability seeks to identify the internal computational structures through which neural networks represent information and generate outputs. Early studies demonstrated that particular neurons in transformer models such as BERT (3) are associated with specific factual knowledge, and that selectively suppressing these neurons can impair factual recall while largely preserving general language capabilities (10). Subsequent research showed that factual associations can be localized and modified within transformer feed-forward networks (11), supporting their interpretation as key-value memories encoding learned relationships (12).

Related investigations have identified neural components associated with linguistic relationships (13), comparative reasoning (14), and contextual processing. These findings contributed to circuit-based approaches to mechanistic interpretability (15), supported by mathematical frameworks describing interactions among transformer components (16). Importantly, distinguishing neural correlates from functionally consequential mechanisms requires causal interventions, including techniques that modify or suppress individual neurons, attention heads, or their contributions to model behavior (17).

A major challenge is that individual neurons need not correspond to unique, independently functioning features. Sparse autoencoders provide a means of decomposing complex activation patterns into more interpretable components (18), addressing the phenomenon of superposition, in which multiple features are represented through overlapping neural dimensions (19). Subsequent developments have improved feature disentanglement (20) and extended these methods to substantially larger language models (21). Complementary visualization techniques, including t-distributed stochastic neighbor embedding (t-SNE) (22) and Uniform Manifold Approximation and Projection (UMAP) (23), facilitate exploration of high-dimensional activation structures,

although spatial separation in these projections does not itself establish functional independence or causality.

Compared with factual knowledge and linguistic processing, the internal representation of emotions has received less systematic investigation. Earlier research examined sentiment processing in recurrent architectures such as long short-term memory networks (24), while more recent work identified a linear representation of sentiment valence in transformer models and demonstrated its causal relevance through targeted interventions (25). The latter provides an important methodological precedent but addresses sentiment polarity rather than the relationships among multiple categorical emotions. Unlike discrete factual associations, emotional representations may overlap, vary with linguistic context and intensity, and participate in several related functions.

2.3 FFN gating and the SwiGLU mechanism

Modern transformer-based language models build on the architecture introduced by Vaswani et al. (26), incorporating feed-forward networks (FFNs) that transform intermediate representations at each transformer layer. Many contemporary architectures, including LLaMA (29), Qwen (30), and Mistral (31), employ the SwiGLU mechanism, which combines gated linear units (27) with the Swish activation function (28). The use of gating to modulate information flow predates these architectures, with earlier applications in gated convolutional neural networks (32).

In a SwiGLU-based FFN, the input is projected through two parallel pathways: one produces a nonlinear gating signal, while the other produces an intermediate feature representation. Their element-wise product determines the intermediate activations passed to a subsequent output projection. Unlike a binary switch, this mechanism continuously modulates information through both the magnitude and sign of the resulting activations. Consequently, FFN gating provides a computational substrate through which different input features can contribute selectively to downstream representations.

These properties make SwiGLU activations a suitable target for investigating emotion-associated neural organization. Examining this activation space may reveal how representations of different emotions overlap and change with task demands.

2.4 Emotion regulation as a source of task contrasts

Psychological theories of emotion regulation distinguish between strategies that modify the appraisal or interpretation of an emotionally significant situation and those that regulate its outward expression. The latter, commonly termed *expressive suppression*, involves inhibiting the visible expression of emotion without necessarily altering the underlying emotional response (33). This distinction provides a useful conceptual framework for examining whether changes in the emotional tone of language-model outputs correspond to changes in their internal emotion-associated representations.

In language models, instructions to express or suppress an emotion can produce different linguistic responses to the same underlying situation. This distinction motivates the fourth experimental task family, which compares emotion-associated activations during expression and suppression, both before generation and in the generated text.

3. Methods

3.1 Dataset description and preprocessing

The study used a bilingual corpus initially comprising 16,667 short Chinese-language texts annotated with primary emotion labels corresponding to Ekman's six basic emotions: anger, disgust, fear, happiness, sadness, and surprise. The texts represented everyday communicative contexts, including product reviews, opinions about films and entertainment, weather-related observations, and conversational expressions. Each Chinese text was accompanied by an English translation generated through an automated translation pipeline. ~ 30% of the samples also contained secondary emotion labels, such as *frustrated*, *anxious*, or *content*. These secondary labels were excluded from the principal analyses and used only in the supplementary independent-annotation assessment (Supplementary Section S3).

Preprocessing involved removing duplicate records identified by identical Chinese text and primary emotion label, excluding records with missing text or unrecognized emotion labels, and removing bracket-wrapping artifacts introduced during translation. Deduplication removed 772 of the 16,667 records, leaving 15,895. No remaining record had missing text or an unrecognized emotion label, so no further records were excluded, and the final analytical corpus comprised 15,895 unique samples, the total of the six class counts below. Thirty-six Chinese texts occurred with more than one primary label (two labels for 30 texts, three for 3 and four for 3; 81 records in all); because duplicates were defined by text and label together, these were retained as separate records.

The distribution of primary emotion labels in the final corpus was unequal: anger accounted for 4,792 samples (30.1%), surprise for 3,095 (19.5%), happiness for 3,010 (18.9%), fear for 1,900 (12.0%), sadness for 1,808 (11.4%), and disgust for 1,290 (8.1%). This imbalance, particularly the smaller representation of disgust, was considered when interpreting emotion-specific activation frequencies, signature reliability, and classification performance.

The Chinese texts served as the primary dataset for identifying emotion-associated neuronal signatures and characterizing their overlap. The corresponding English translations were used to assess cross-language reproducibility of the baseline representations. The experimental task families described in Section 3.6 used English-language inputs, including held-out corpus texts and independently constructed scenarios, to examine task-dependent representational changes and their relationship to observable model behavior.

The primary emotion labels were treated as the reference annotations for statistical

comparisons, rather than as definitive descriptions of each text's complete emotional content. Consequently, the analysis characterizes associations between model activations and the assigned emotion categories, while recognizing that natural-language expressions may contain multiple or ambiguous emotions.

3.2 Model architecture and selection rationale

The study employed Qwen2.5-1.5B-Instruct, an instruction-tuned transformer language model comprising 28 layers, each incorporating grouped-query attention and a SwiGLU-based feed-forward network (FFN). The model has a hidden-state dimension of 1,536 and an FFN intermediate dimension of 8,960. Model weights were loaded in bfloat16 precision. The initial activation extraction was performed on an NVIDIA RTX 2080 Ti GPU with 11 GB of video memory. Subsequent FFN activation re-extraction and experimental analyses were conducted on a CPU using two threads, PyTorch 2.14.1, and Transformers 5.18.0, while retaining the same model weights and numerical precision.

The selection of a 1.5-billion-parameter model balanced the need to investigate complex internal representations against the computational demands of extracting neuron-level activations from 15,895 texts. This model size permitted systematic analysis of the selected FFN layers across the full corpus without requiring the substantially greater resources associated with larger architectures. However, the use of a single model limits the generalizability of the findings, as discussed in Sections 6 and 7.3.

The analysis focused on nine consecutive transformer layers, 14–22 inclusive. These layers were selected on the premise that intermediate-to-later transformer representations are more likely to contain abstract semantic information than representations dominated by lower-level lexical and syntactic features. Because emotion is a context-dependent semantic property, this region was considered a plausible location for identifying emotion-associated activation patterns. Nevertheless, the layer selection represents a methodological assumption rather than an experimentally established localization of emotional processing. The findings therefore characterize the selected layers and do not exclude emotion-associated representations or relevant computations elsewhere in the model.

3.3 Activation extraction pipeline

Internal activations were extracted from layers 14–22 of Qwen2.5-1.5B-Instruct to characterize emotion-associated representations in the feed-forward networks (FFNs) and residual stream. Two complementary activation spaces were examined: the post-SwiGLU intermediate activations within each FFN and the residual-stream hidden states following each transformer block. The former provided the primary basis for identifying emotion-associated neuronal signatures, while the latter was used to examine the broader geometry of the model's internal representations.

3.3.1 *FFN gate activations (Post-SwiGLU)*

The primary analytical target was the post-SwiGLU intermediate activation vector within each selected FFN layer. PyTorch forward hooks were registered on the multilayer

perceptron (MLP) modules to capture the element-wise product of the activated gate projection and the up-projection, before the result was passed through the down-projection: SiLU(gate_proj(x)) × up_proj(x), where x denotes the layer input and SiLU is the sigmoid linear unit activation, also known as Swish. For each layer, the extracted activations formed a tensor of dimensions [batch_size, sequence_length, 8960], corresponding to 8,960 intermediate FFN dimensions at each token position. These dimensions are referred to operationally as *gating neurons* throughout the study.

Because the recorded quantity is the product of the gate and up-projection pathways, rather than the gate projection alone, the analysis characterizes the combined post-SwiGLU intermediate representation. Associations identified in this space cannot be attributed exclusively to the gating pathway.

3.3.2 *Residual-stream hidden states*

Residual-stream hidden states were also extracted from the output of each selected transformer block. Unlike the post-SwiGLU intermediate activations, these states incorporate the cumulative contributions of attention, normalization, and feed-forward processing within the transformer architecture.

For each layer, the residual-stream activations formed a tensor of dimensions [batch_size, sequence_length, 1536], corresponding to the model's hidden-state dimensionality.

These representations provided a complementary basis for examining emotion-associated structure beyond the FFN intermediate space. Figures 9 and 10 were produced from the original residual-stream extraction and were not regenerated during the corrected FFN signature analysis; Figure 11 was regenerated from the verification re-extraction described below. Subsequently, selected residual-stream analyses were independently repeated to verify the unusually high PCA explained variance and assess the effects of feature standardization and sequence length. The verification results are reported in Section 4.3.

3.3.3 *Token-level aggregation and activation thresholding*

Because input texts varied in length, token-level activations were aggregated using attention-masked mean pooling. This produced one activation vector per text and layer, containing 8,960 dimensions for the FFN intermediate activations and 1,536 dimensions for the residual stream. Masking excluded padding tokens from the pooled representations.

For each sample, the most strongly activated dimensions were identified by ranking their absolute activation magnitudes. A top-K percentage threshold was then applied to convert the continuous activation vector into a binary activation mask. The default threshold was 5%, corresponding to the 448 highest-magnitude of the 8,960 FFN dimensions per sample and layer. Binarization was applied only to the FFN activations; the residual-stream analyses (Section 3.4.3) used the continuous pooled vectors.

The threshold was defined relative to each

sample's activation distribution rather than by a fixed activation magnitude. Consequently, an *active neuron* denotes a dimension selected by this operational criterion, not necessarily one that is functionally engaged or required for the model's response.

Sensitivity analyses were conducted at thresholds of 1%, 3%, 5%, and 10% to evaluate the dependence of the findings on the activation-selection criterion. FFN intermediate activations were extracted for all 15,895 texts in both their original Chinese form and their English translations. Texts were processed in batches of 32, sorted by length, with right padding, truncation at 512 tokens, and attention-masked mean pooling. The pooled activation arrays were stored in float16 format.

To assess extraction reproducibility, the newly extracted activations were compared with the earlier uncorrected results. Signature-neuron totals differed by no more than 0.2% per emotion, while the 15 pairwise overlap values differed by no more than 0.0016. These small discrepancies were attributed to differences in batch composition and padding. The corresponding comparisons are reported in Supplementary Table S1.

The resulting pooled activation vectors and binary masks formed the basis for the statistical identification of emotion-associated neuronal signatures, overlap analyses, and cross-language comparisons described in Section 3.4.

3.3.4 *Implementation and reproducibility*

The commercial pipeline used Python, Pytorch Hugging Face Transformers, SciPy, scikit-learn, and statsmodels (36), with figures generated in Matplotlib at 300 DPI. In the original study, the complete pipeline, comprising activation extraction for both activation spaces, all statistical analyses and figure generation, required approximately 14 hours on an NVIDIA RTX 2080 Ti GPU (11 GB) with batches of eight. This total was recorded for the pipeline as a whole, and the extraction step was not timed separately, so it is not comparable with the extraction time reported below. For texts of this length (median 17 tokens and maximum 70 for the Chinese texts, so that no text reached the 512-token truncation limit), one forward pass per text is a small computational load, whereas the original analyses included 1,000-permutation tests, 100 split-half repetitions per emotion, five-fold cross-validated logistic regression on 80,640 binary features at each of four thresholds, and t-SNE and UMAP embeddings of 15,895 vectors of 13,824 dimensions. In the present revision each step was timed separately on a CPU with two threads: FFN activation extraction took approximately 25 minutes per language corpus (batches of 32 sorted by length), the corrected baseline analyses approximately 20 minutes per language, the nested cross-validated classifier approximately 2.1 hours, the verification re-extraction of the residual stream approximately 14 minutes, and the four experimental task families approximately three hours in total (all four were run between the preregistration freeze at 20:36 and 23:40 UTC on 5 October 2026; Supplementary Section S11). The computational procedures, experimental stimuli, preregistration materials, and trial-level

results are identified in the Data and Code Availability statement.

3.4 Identification and validation of emotion-associated representations

The following analyses identify and validate emotion-associated signatures, quantify their overlap, and assess their correspondence with psychological emotion structure, lexical similarity, and activation-based classification errors.

3.4.1 *Statistical analysis of emotion-associated neuronal activation*

For each emotion, neuronal activation frequency was calculated as the proportion of samples in which a neuron was classified as active. Neurons activated in more than 50% of samples belonging to a particular emotion were designated consistently activated neurons.

χ^2 tests of independence assessed associations between neuronal activation and emotion labels. The resulting p-values were corrected for multiple comparisons using the Benjamini–Hochberg procedure (34), separately within each layer and emotion. Both uncorrected and corrected counts were retained for comparison. Additionally, 1,000 label-permutation iterations were performed to establish the null distribution of consistently activated neuron counts under random emotion-label assignment.

These analyses determine whether emotion-associated activation frequencies exhibit statistically distinguishable structure, without assuming that the identified neurons are exclusively associated with particular emotions or causally responsible for emotional behavior.

3.4.2 *Identification and reliability assessment of emotion signatures*

Emotion-associated signature neurons were identified using one-sided two-proportion z-tests comparing each neuron's activation frequency in samples of a target emotion with its frequency in all other samples.

The original selection criterion of an uncorrected p-value below 0.05 was replaced because testing 8,960 neurons per layer and emotion introduces substantial multiple-comparison risk. Under the corrected criterion, a neuron qualified as an emotion-associated signature neuron only when its Benjamini–Hochberg-adjusted p-value was below 0.05 within the corresponding family of 8,960 tests (34), its effect size was at least 0.2 as measured by Cohen's h (35), and its activation frequency was higher for the target emotion than for the remaining samples.

Cohen's h was calculated as: $h = 2\arcsin(\sqrt{p_t}) - 2\arcsin(\sqrt{p_o})$, where p_t represents the activation frequency in the target-emotion samples and p_o represents the frequency in all other samples. The effect-size threshold excludes associations that are statistically significant but negligible in magnitude.

Signature reliability was evaluated through repeated split-half analysis. Samples belonging to each emotion were randomly divided into two halves, signature sets were independently recomputed, and their Jaccard similarity was calculated. This procedure was repeated 100 times at layer 18.

A complementary five-fold stratified cross-

validated logistic regression classifier was trained on concatenated binary FFN activation masks from the nine target layers. Because the classifier used label-independent activation features, no signature-selection procedure was required for its basic cross-validation. However, whenever signature overlap was compared with held-out classifier predictions, signature identification was nested within the training folds to prevent information leakage.

All analyses dependent on signature selection were repeated using the corrected signatures. The original uncorrected results documenting the consequences of the revised selection criterion were retained in Supplementary Tables S1 and S2.

3.4.3 *Residual-stream dimensionality reduction and clustering analysis*

The organization of emotion-associated representations was further investigated using residual-stream activations concatenated across the nine target layers, producing a 13,824-dimensional feature vector for each sample.

Principal component analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE), and Uniform Manifold Approximation and Projection (UMAP) were used to obtain two-dimensional visualizations. The input to each was the residual-stream vector of each text, mean-pooled over its tokens and concatenated across layers 14 to 22 (9 x 1,536 = 13,824 dimensions). PCA was computed with scikit-learn, which centres each dimension but does not scale it, and the explained variance of a component was its eigenvalue divided by the total variance. For t-SNE, the original feature vectors were first reduced to 50 dimensions using PCA. The projections were colored according to emotion labels to examine whether samples associated with different emotions occupied distinguishable regions. Silhouette analysis quantified cluster separation, while k-nearest-neighbor purity, with k = 15 and cosine distance, measured local consistency of emotion labels. Purity was computed in the 50-component PCA space, with each sample excluded from its own neighbors, and was compared for each emotion with that emotion's prevalence in the corpus, which is the purity expected if neighbors are unrelated to the label.

These procedures characterize the geometry of the residual-stream representations. Visual clustering does not, by itself, establish that emotions occupy discrete functional modules or that the observed dimensions govern model behavior.

3.4.4 *Quantification of overlap between emotion signatures*

To characterize relationships among emotion-associated signatures, pairwise Jaccard similarities were calculated between the corrected signature neuron sets for all six emotions. Each neuron was represented by a unique (layer, neuron index) tuple, permitting aggregation across layers without conflating neurons occupying identical indices at different depths.

Neurons associated exclusively with one emotion were distinguished from those shared by multiple emotions. Hierarchical clustering using average linkage on the Jaccard distance matrix was performed to construct a

dendrogram of emotion similarity. An UpSet-style intersection analysis quantified the sizes of the distinct shared and exclusive signature groups. These analyses determine whether emotion-associated neuronal signatures are largely distinct or exhibit systematic overlap. Overlap is interpreted as shared statistical association, rather than evidence that individual neurons independently encode or control multiple emotions.

3.4.5 *Testing correspondence between neuronal overlap and psychological emotion structure*

To determine whether the observed neuronal overlap corresponds to established psychological relationships among emotions, pairwise Jaccard similarities were compared with two matrices derived from the circumplex model of affect.

The six emotions were organized into three proposed groups: happiness and surprise; anger and disgust; and fear and sadness. The first matrix was binary, assigning a value of one to emotion pairs within the same group and zero otherwise. The second was graded, assigning two to within-group pairs, one to pairs sharing valence but belonging to different groups, and zero to pairs with opposing valence.

Spearman rank correlations were calculated between the 15 observed pairwise Jaccard similarities and the corresponding values in each model matrix. Statistical significance was assessed through exact permutation tests encompassing all 720 possible relabelings of the six emotions. The cophenetic correlation coefficient was also calculated to assess how faithfully the hierarchical clustering dendrogram represented the underlying Jaccard distance matrix. These tests provide a quantitative assessment of psychological correspondence rather than relying on subjective visual resemblance between the neuronal overlap structure and the circumplex model.

3.4.6 *Assessing the influence of lexical similarity on emotion-signature overlap*

An apparent correspondence between neuronal overlap and psychological emotion categories could reflect similarities in the language used to express those emotions. To investigate this possibility, the English translations of the corpus were aggregated into six documents, one for each emotion class.

Term-frequency–inverse-document-frequency (TF-IDF) representations were constructed after removing English stop words, applying sublinear term-frequency scaling, and restricting the vocabulary to 20,000 terms. Pairwise cosine similarities between the resulting class vectors produced a lexical similarity matrix directly comparable with the neuronal overlap matrix.

Three associations were evaluated: the Spearman correlation between neuronal overlap and the circumplex model, the correlation between lexical similarity and the circumplex model, and the partial Spearman correlation between neuronal overlap and the circumplex model after controlling for lexical similarity.

The significance of the partial correlation was assessed using exact permutation over all 720 possible emotion-label relabelings. The lexical

similarity and neuronal overlap matrices were permuted jointly to preserve their dependence under the null hypothesis. This analysis tests whether the observed psychological correspondence is attributable to the specific measure of surface lexical similarity. It does not exclude other linguistic or semantic confounders.

3.4.7 *Consistency of emotion-associated activation patterns*

The final baseline analysis examined whether aggregate overlap among emotion-associated neuronal signatures corresponded to the confusability of individual samples within the same FFN activation space.

A logistic regression classifier was evaluated using five-fold stratified cross-validation on binary FFN activation masks. To prevent information leakage, corrected signature sets were recomputed using only the training samples in each fold. The resulting signature-overlap matrix was then compared with the classifier's confusion pattern on the corresponding held-out samples and the mean signature-overlap matrix across folds reported.

Confusion-matrix entries were converted into lift relative to the marginal prediction rate, logarithmically transformed, and symmetrized. Spearman correlation across the 15 emotion pairs quantified agreement between signature overlap and activation-based confusability. Statistical significance was determined through exact permutation over all 720 possible emotion-label relabelings. Because both the classifier and the signature sets were derived from FFN activation patterns, this comparison measures internal activation consistency rather than independent model behavior. Even a strong correspondence would not establish that the model uses the identified signatures when generating emotional responses.

3.4.8 *Testing the portability of emotion-associated neuron indices*

To assess whether emotion-associated neuron indices remain consistent across layers, the recurrence of corrected signature neuron indices was examined between different layers of the same model for each emotion. Observed overlap was compared with the expected overlap under a hypergeometric null model, assuming random selection of neuron sets of equivalent sizes from the 8,960 available indices.

The same analysis was performed across different emotions within a single layer to provide a comparison with cross-layer recurrence. This distinction allowed the analysis to examine whether particular neuron indices recur more frequently than expected by chance, either across layers for the same emotion or across emotions within a layer.

Because neuron indices in different layers do not necessarily correspond to functionally equivalent units, cross-layer recurrence was interpreted as a statistical property rather than evidence of conserved neuronal identity or function.

3.4.9 *Cross-language replication (RQ1: Representation)*

To assess the robustness of emotion-associated representations across languages, the complete baseline analysis was repeated using FFN gating

activations extracted from English translations of the original Chinese texts. The same activation-processing and corrected signature-selection procedures were applied independently to both language versions.

Cross-language agreement was evaluated using two complementary measures. First, Spearman rank correlation was calculated between the Chinese and English corrected signature-overlap matrices across all 15 emotion pairs. Statistical significance was assessed through exact permutation over all 720 possible relabelings of the six emotions.

Second, for each emotion, Jaccard similarity was calculated between its Chinese and English signature neuron sets. Observed similarities were compared with the overlap expected from randomly selected sets of equivalent sizes drawn from the 80,640 possible layer–neuron positions.

The two analyses separately assessed the reproducibility of emotion-overlap structure and the recurrence of individual signature neurons. Because the comparison used translations of the same underlying texts within a single model, it did not establish generalizability to independent corpora or architectures.

3.5 Robustness and sensitivity analyses

To assess sensitivity to the activation-selection criterion defined in Section 3.3.3, the analyses were repeated at thresholds of 1%, 3%, 5%, and 10%. At each threshold, four quantities were recalculated: the number of consistently activated neurons, the number of emotion-associated signature neurons satisfying the corrected statistical and effect-size criteria, the pairwise Jaccard overlap matrix for the six emotions, and the Spearman correlation between neuronal overlap and the graded circumplex model. The primary objective was to determine whether the relative organization of emotion-associated overlap and its correspondence with the circumplex model remained stable despite changes in the number and identity of selected neurons.

3.6 Experiments on task-dependent reconfiguration of emotion signatures

3.6.1 *Preregistration and independence of experimental stimuli*

The four experimental task families were specified in a preregistration document before extraction of any experimental activations (37). The preregistration defined the hypotheses and their predicted directions, experimental stimuli, internal activation and behavioral measures, exclusion criteria, statistical models, and multiple-comparison correction procedures. This prospective specification was intended to distinguish confirmatory hypothesis testing from analyses developed after examining the experimental results.

To document the timing and integrity of the preregistration, SHA-256 hashes were generated for the preregistration document, stimulus files, reference map, and experimental and analytical code. These hashes were recorded with a coordinated universal time (UTC) timestamp of 5 October 2026 at 20:36:36, before the experiments commenced. Although the preregistration was not deposited in a public registry, the document and corresponding hash

records are provided in the supplementary materials to permit verification of subsequent alterations. One hashed file, the analysis script, was modified after the freeze (Section 5.1).

Before the preregistration was finalized, a pilot run using no study stimuli was conducted solely to verify pipeline operation and estimate generation time. The analytical code was also tested using synthetic datasets. Under simulated null conditions, none of the confirmatory tests remained significant after multiple-comparison correction; when effects corresponding to the preregistered hypotheses were introduced, all were detected. These checks assessed the implementation of the analytical procedures but did not constitute validation of the experimental hypotheses.

To reduce dependence between the baseline neuronal reference map and its subsequent experimental evaluation, the English-language corpus was divided once into a stratified derivation set of 11,126 texts (70%) and a held-out set of 4,769 texts (30%). The reference map used in the reconfiguration experiments was constructed exclusively from the derivation set. Experimental Family 1 used only held-out corpus texts. Families 2–4 used newly constructed stimuli that were not included in the original corpus. The provenance and preparation of these stimuli, including any use of generative artificial intelligence, are described in Section 3.6.3.

3.6.2 *Reference emotion map and quantification of internal reconfiguration*

The reference map comprised the corrected signature neuron sets for the six emotions across transformer layers 14–22, derived exclusively from the English-language derivation set using the criteria described in Section 3.4.2. The map also included the mean and standard deviation of each neuron's pooled activation. These reference values were used to standardize neuronal activations during the subsequent experiments. The resulting signature sets contained 567 neurons for anger, 126 for disgust, 592 for fear, 677 for happiness, 320 for sadness, and 1,334 for surprise.

Three internal measures were used and are kept distinct throughout. For a given measurement position, let x_{lj} denote the activation of neuron j in layer l, averaged over the tokens of that position; let Se denote the set of (layer, neuron) pairs in the signature of emotion e, pooled across the nine layers, with $|Se|$ members; and let μ_{lj} and σ_{lj} denote the mean and standard deviation of the pooled activation of that neuron in the derivation set.

[1] Preregistered measure (signed standardized activation): $se = (1/|Se|) \Sigma (x_{lj} - \mu_{lj}) / \sigma_{lj}$, where here and below the sum runs over all (l, j) in Se. This is the confirmatory measure specified in the preregistration. Because it is linear, averaging token-level scores over a span is identical to scoring the span-averaged activation.

[2] Primary corrected measure (mask fraction): $se = (1/|Se|) \Sigma 1[\, |x_{lj}| \geq \tau_l \,]$, where τ_l is the 448th largest absolute value among the 8,960 pooled activations of layer l at that position (the top 5%, the binarization used to define the signatures in Section 3.4.2) and $1[\cdot]$ equals 1 when the condition holds and 0 otherwise. The score is the fraction of the signature neurons of emotion e that are active.

[3] Robustness corrected measure (absolute standardized activation): $se = (1/|Se|) \Sigma (|xlj| - \mu^{abs}lj) / \sigma^{abs}lj$, where $\mu^{abs}lj$ and $\sigma^{abs}lj$ are the mean and standard deviation of the absolute pooled activation in the derivation set.

For each measure, the pairwise activation axis for emotions *a* and *b* is $Aab = sa - sb$. Measure [1] was fixed before the experiments. Measures [2] and [3] were introduced afterwards, through the documented deviation described in Section 5.1, because measure [1] averages signed activations whereas the signatures are defined by absolute activation; results under measures [2] and [3] are therefore *post hoc* analyses.

The three measures quantify activation within emotion-associated signature sets established from the baseline derivation data. The signature sets were fixed independently of the experimental outcomes. The signed standardized activation measure was preregistered, whereas the mask-fraction and absolute-standardized-activation measures were introduced after the preregistered results of Families 1–3 had been examined (5.1). The corrected measures were selected using held-out baseline data and subsequently applied to the saved experimental activations. Accordingly, they are post hoc measures rather than prospectively specified confirmatory measures. None of the three scores represents a calibrated probability of an emotional state or direct evidence of neuronal function.

The preregistration specified three locations for measuring internal activations, depending on the experimental task: the tokens comprising the stimulus text, the final prompt token immediately preceding response generation, and the tokens of the generated response. Because the task instruction preceded the stimulus, stimulus-token activations were measured in the context of the instruction rather than independent of it. These measurement locations permitted comparisons of emotion-associated activation during stimulus processing, immediately before response generation, and during the generated response.

The reference overlap matrix derived from the English-language derivation set showed the highest Jaccard similarities for anger–disgust (0.068) and fear–sadness (0.054), followed by disgust–sadness (0.028). Happiness–surprise, which had been among the higher-overlap pairs in the full Chinese-language baseline analysis, exhibited substantially lower overlap in the English derivation set (0.004).

This discrepancy was documented in the preregistration before the experiments. The happiness–surprise pair was retained in the experimental design, and results were reported separately for each emotion pair rather than modifying the original comparisons in response to the observed overlap values.

3.6.3 *Four experimental families for testing task-dependent reconfiguration*

All experimental prompts used the model's chat template with a single user turn, in which the task instruction preceded the stimulus text. Four paraphrases were constructed for each instruction. For forced-choice tasks, the order of the two response options was reversed in odd-numbered paraphrases. The four preregistered task families are summarized in Table 1.

Table 1. Design of the four preregistered task families.

Family	Items	Conditions	Trials	Internal measure	Behavioural measure
1. Recognition vs discrimination	480 held-out texts (160 per pair)	RECOGNIZE, DISCRIMINATE	3,840	Pair axis over stimulus tokens, signed to gold	Log-probability of gold minus other label
2. Disambiguation after evidence	120 written scenarios (40 per pair)	NONE, FILLER, EV_A, EV_B	1,920	Pair axis at final token	Log-probability of label a minus label b
3. Targeted tone generation	40 neutral statements x 6 targets	Target emotion	960	Centred target margin over generated tokens	Emotion classifier; NRC lexicon
4. Expression vs suppression	40 anger and 40 fear situations	EXPRESS, SUPPRESS, NEUTRAL	960	Target score over generated tokens and at planning state	Classifier probability of target

Family 1 used human-written texts from the held-out split of the corpus, with their human-assigned primary labels. The stimuli of Families 2 to 4 were drafted with a generative AI system (Claude, Anthropic) under the direction of the authors, to the specifications given below for each family: the 120 Family 2 scenarios with their evidence and filler sentences, the 40 neutral Family 3 statements, the 80 Family 4 situations, and the four paraphrases of each task instruction. Every item was reviewed by the authors. The Family 2 evidence sentences were checked by regular expression for the six emotion-label words and their direct morphological variants (0 violations). All stimulus files were fixed and hashed before the preregistration was frozen (Section 3.6.1), and the complete sets are released in the supplementary archive. Because these stimuli were machine-drafted rather than written by independent human authors or drawn from a corpus, their language may differ systematically from naturally occurring text (Section 7.3).

3.6.3.1 Family 1: Recognition versus forced discrimination

Three emotion pairs were examined: anger–disgust, fear–sadness, and happiness–surprise. For each pair, 80 held-out texts were sampled from each constituent emotion, yielding 480 texts of 6–60 words. The original human-assigned primary emotion served as the reference label.

Each text was presented under two conditions: RECOGNISE, in which the model was instructed to identify every emotion potentially present, and DISCRIMINATE, in which it was required to choose between the two emotions in the corresponding pair. The internal measure was the pairwise activation axis averaged across the stimulus tokens and signed toward the reference label.

The preregistered prediction was that forced discrimination would produce greater separation between the paired emotions than broad recognition.

3.6.3.2 Family 2: Disambiguation following additional evidence

For each of the three emotion pairs, 40 scenarios were constructed to be initially compatible with either emotion, yielding 120 scenarios. Each scenario was accompanied by two alternative evidence sentences, one favoring each emotion,

and a neutral filler sentence. The evidence sentences excluded the emotion-label words and their direct morphological variants. Instead, the evidence was conveyed through descriptions of behavior, appraisal, physiological responses, or speech. For example, anger could be suggested by someone slamming a refrigerator door and demanding to know who had taken their food, whereas disgust could be suggested by someone gagging and pushing away a moldy container.

Each scenario was presented under four conditions: NONE (scenario alone), FILLER (scenario with the neutral sentence), EV_A (evidence favoring the first emotion), and EV_B (evidence favoring the second emotion). In every condition, the model was instructed to choose between the two emotions.

The internal measure was the pairwise activation axis at the final prompt token. The preregistered prediction was that the axis would shift toward the emotion favored by the additional evidence, with a greater shift than that produced by adding an irrelevant sentence.

3.6.3.3 Family 3: Targeted emotional tone generation

Forty emotionally neutral factual statements were written, such as a statement indicating that a meeting had been rescheduled from Monday to Thursday. The model was instructed to rewrite each statement to express each of the six target emotions while preserving its factual content. The experiment comprised 960 generations.

The internal measure was the target emotion score across the generated tokens. This score was obtained through a forward pass over the combined prompt and generated response and centered within each statement and instruction paraphrase across the six target emotions. Two predictions were preregistered: first, that the target emotion's activation score would exceed the scores of the other five emotions; and second, that emotion pairs with greater overlap in the reference map would be more difficult to distinguish, both internally and behaviorally.

3.6.3.4 Family 4: Emotional expression versus suppression

Eighty situations were constructed, comprising 40 associated with anger and 40 associated with fear. Examples included a colleague taking credit for one's design (anger) and receiving threatening calls from an unknown number (fear).

For each situation, the model was instructed to compose a message under three conditions: EXPRESS, in which the relevant emotion was communicated openly; SUPPRESS, in which the same concern was communicated while withholding overt emotional expression; and NEUTRAL, in which a factual report was requested. The experiment comprised 960 generations.

Two internal measures were obtained: the target emotion score across the generated tokens and the corresponding score at the final prompt token before generation, designated the *planning state*.

The preregistered prediction was that the target emotion score would be higher under EXPRESS than SUPPRESS in the generated text. It was

further predicted that this difference would be greater in the generated text than at the planning state, consistent with suppression operating primarily at the level of output rather than through a corresponding change in the preceding internal state.

3.6.4 *Independent measures of behavioral performance*

For Families 1 and 2, the assistant response was prefixed with "Answer:", and the next-token log-probabilities of the six emotion-label words were extracted. Each label corresponded to a single token in the model's tokenizer. The behavioral margin was calculated as the difference between the log-probabilities of the two emotion labels being compared. A trial was classified as correct when the margin favored the reference emotion in Family 1 or the emotion supported by the additional evidence in Family 2.

For Families 3 and 4, responses were generated using greedy decoding, with a maximum of 80 new tokens per response. The generated texts were evaluated using the independent emotion classifier *emotion-english-distilroberta-base* (38), which assigns probabilities to seven categories: anger, disgust, fear, joy, neutral, sadness, and surprise. The classifier's joy category was mapped to happiness to maintain consistency with the six emotion categories used in the study. As a second behavioral measure, the NRC Emotion Lexicon (1) was used to calculate the proportion of words associated with each emotion in the generated responses.

3.6.5 *Statistical testing of task effects and behavioral associations*

Because all four experimental families included repeated measurements across items and instruction paraphrases, items and paraphrases were treated as crossed random factors (39, 40). Continuous outcomes were analyzed using linear mixed-effects models with crossed random intercepts for item and paraphrase, fitted by restricted maximum likelihood. Binary outcomes were analyzed using corresponding logistic mixed-effects models fitted by variational Bayes (36).

Fitting linear mixed-effects models with several hundred item levels was computationally impractical using the statsmodels MixedLM implementation specified in the preregistration. The linear models were therefore fitted with a purpose-written Python implementation (lmm_crossed.py, released with the code) of the penalized least-squares formulation of restricted maximum likelihood used in lme4 (41); the R package lme4 itself was not used. The model, estimator and test are those of the preregistration. On synthetic data, this implementation reproduced the statsmodels MixedLM fixed-effect estimate and its standard error to six significant figures and the variance components to five (Supplementary Section S6). All linear mixed-model results in Table 3 and Supplementary Tables S7 and S8 were produced with this implementation, and the logistic models were fitted with statsmodels BinomialBayesMixedGLM, as preregistered. The change of fitting software is Deviation 2 (Section 5.1; Supplementary Section S7.1).

Where specified in the preregistration, nonparametric tests were used. In Family 1, the null distribution for the change in paraphrase-

level d-prime was constructed by exhaustively exchanging condition labels among the eight paraphrase-condition cells for each emotion pair, yielding 343,000 combinations. This procedure provided an empirical null distribution reflecting variation attributable to prompt wording. In Family 2, a Wilcoxon signed-rank test compared the activation shift associated with emotion-specific evidence against the shift produced by the neutral filler. In Family 3, the association between pairwise emotion separability and reference-map overlap was evaluated using an exact permutation test encompassing all 720 possible relabelings of the six emotions.

Fourteen confirmatory hypotheses were preregistered, each specifying a one-sided directional prediction (Table 3). The family-wise error rate across these hypotheses was controlled at 0.05 using the Holm multiple-comparison procedure (42). The Family 4 manipulation check and all secondary analyses were excluded from this confirmatory correction family. These secondary analyses included estimates for individual emotion pairs and instruction paraphrases, replication analyses using odd- and even-numbered item halves, and evaluations using the NRC Emotion Lexicon. The sole preregistered exclusion criterion was an empty generated response.

4. Results: The corrected baseline

All values in this section are computed with the corrected signature criterion of Section 3.4.2 on the re-extracted Chinese activations unless stated otherwise, and the English replication is reported alongside. The values published in the earlier version, the re-extracted uncorrected values and the corrected values are set side by side in Supplementary Tables S1 and S2, and the principal statistics are summarised in Table 2.

Table 2. Principal baseline statistics in the earlier version and after correction (BH-FDR q < 0.05 within each layer and emotion, Cohen's h of at least 0.2). FS: fear and sadness; AD: anger and disgust; HS: happiness and surprise; DS: disgust and sadness. Exact p-values are from 720 relabelings (floor 0.0014).

Statistic	Earlier version (uncorrected)	Corrected, Chinese	Corrected, English
Signature neurons, all emotions	96,743	4,444	3,569
Three highest-overlap pairs	FS, AD, HS	FS, AD, HS	AD, FS, DS
Graded circumplex ρ (exact p)	0.861 (0.022)	0.888 (0.022)	0.773 (0.022)
Binary circumplex ρ (exact p)	0.694 (0.067)	0.697 (0.067)	0.550 (0.067)
Partial ρ given lexical (exact p)	0.868 (0.022)	0.855 (0.022)	0.714 (0.022)
Cophenetic correlation	0.948	0.990	0.952
Overlap vs classifier confusability, ρ (exact p)	0.714 (0.0042), not nested	0.789 (0.0014), nested	not computed
Chinese vs English overlap matrices, ρ (exact p)	not computed	0.920 (0.0014)	

4.1 Emotion-dependent neuronal activation

4.1.1 *Consistent neuron frequency across layers*

The first analysis examined how many FFN gate neurons are consistently activated (frequency > 50%) for each emotion at each layer (Figure 1).

This count does not depend on signature selection. It rises from layer 14 (78 to 90 neurons per emotion) to layer 22 (251 to 286). Anger is at or near the highest count at every layer, and happiness and surprise are generally the lowest.

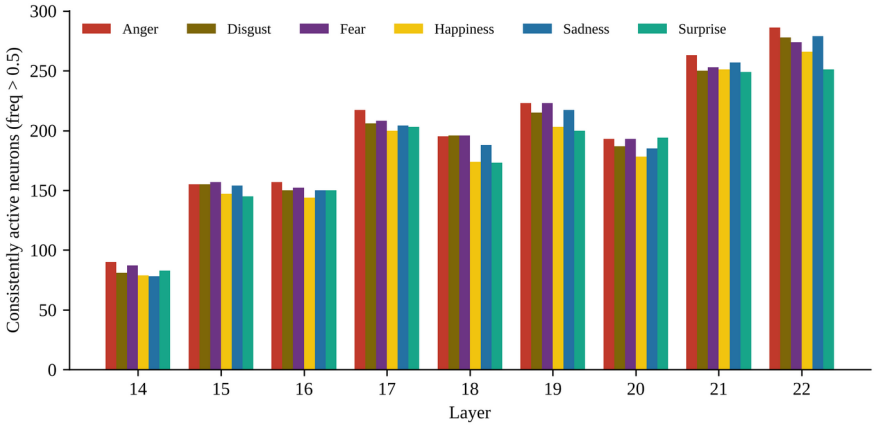


Figure 1. Number of consistently activated FFN gate neurons (frequency > 0.5) per emotion at each layer, Chinese originals. Anger is at or near the top at every layer, and happiness and surprise are generally lowest.

4.1.2 *Chi-Square tests for neuron-emotion association*

Without correction, surprise and anger show the largest numbers of neurons associated with the emotion label at layer 14 (4,396 and 4,157) and disgust the fewest (1,485). After Benjamini-Hochberg correction within each layer and emotion, most associations survive for the larger classes (3,529 for anger and 3,871 for surprise at layer 14), but only about a third survive for disgust (500 at layer 14), the class with the fewest samples (Figure 2). The ordering of emotions and the decline in counts from earlier to later layers are unchanged by the correction.

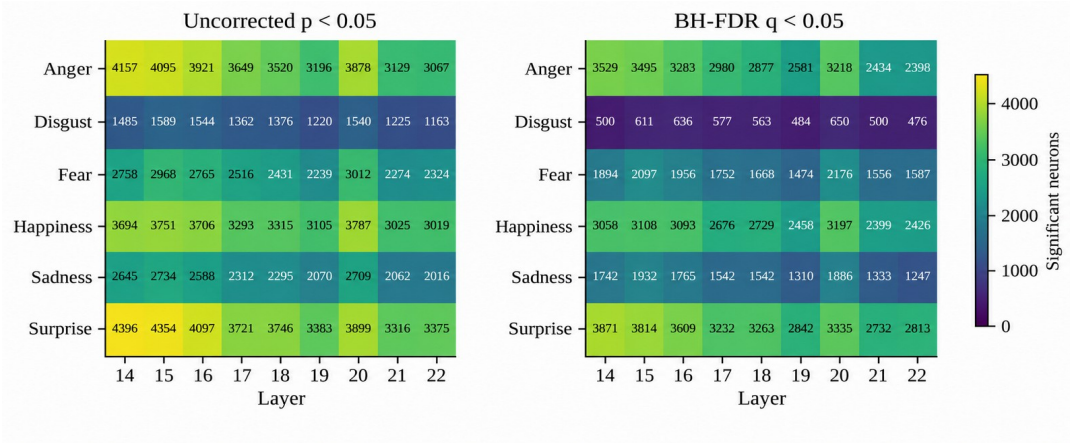


Figure 2. Number of FFN neurons whose activation is associated with each emotion (2 x 2 χ^2 test), per layer. Left: uncorrected $p < 0.05$, as in the earlier version. Right: Benjamini-Hochberg $q < 0.05$ within each layer and emotion.

4.1.3 *Permutation tests of emotion-dependent activation frequencies*

At layer 18, the observed number of consistently activated neurons was compared with null distributions generated from 1,000 emotion-label permutations. Counts exceeded the permutation null for anger (195 versus a null mean of 178.6; p = 0.0010), disgust (196 versus 179.9; p = 0.0010), fear (196 versus 179.5; p = 0.0010), and sadness (188 versus 179.3; p = 0.002). Happiness (174; p = 0.988) and surprise (173; p = 0.995) showed no corresponding excess (Figure 3). English activations showed similar results for anger, disgust, fear, and sadness (all p ≤ 0.002) and surprise (p = 0.990), while happiness was not significant (p = 0.265).

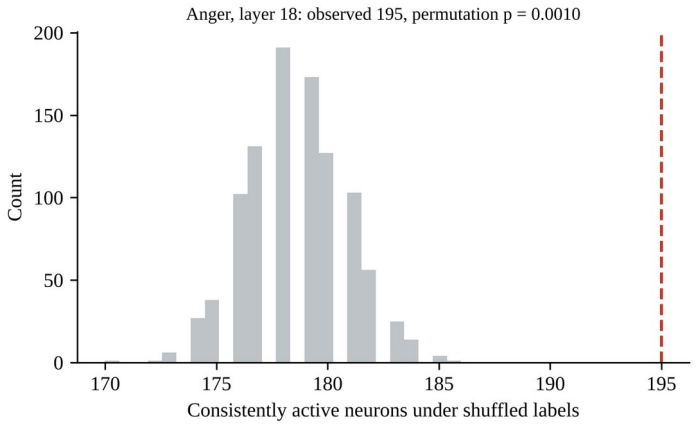


Figure 3. Permutation test for anger at layer 18 (Chinese originals). The observed count of consistently activated neurons (dashed line) lies beyond all 1,000 permuted values.

4.1.4 *Volcano plot analysis: effect size and statistical significance*

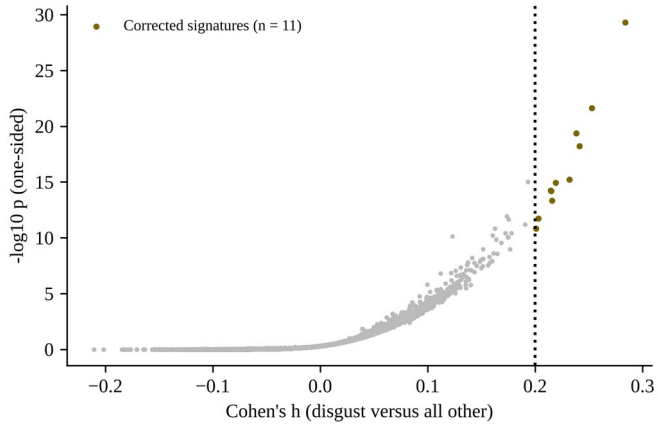


Figure 4. Volcano plot for disgust at layer 14 (Chinese originals). Coloured points are the corrected signature neurons (BH-FDR q < 0.05 and Cohen's h of at least 0.2, dotted line).

The volcano plot for disgust at layer 14 (Figure 4) plots neuronal effect sizes (Cohen's h) against statistical significance. Under the original uncorrected signature-selection criterion based on a one-sided two-proportion z-test, 1,161 neurons qualified, compared with 11 under the corrected criterion requiring both BH-FDR $q < 0.05$ and Cohen's $h \geq 0.2$. The uncorrected count of 1,161 is different from the 1,485 neurons identified by the separate χ^2 association test in Section 4.1.2, which uses a different statistical criterion. The correction excluded numerous small-effect associations, while the small corrected disgust signature set is also associated with low split-half reliability (Section 4.2.2).

4.2 Identification and reliability of emotion signatures

4.2.1 *Signature neuron identification under multiple-testing correction*

Across layers 14–22, the corrected Chinese signature sets contained 844 neurons for anger, 106 for disgust, 818 for fear, 739 for happiness, 473 for sadness, and 1,464 for surprise, totaling 4,444 (Figure 5). Surprise had the largest set and disgust the smallest. Surprise signatures peaked at layers 17 and 18 (201 and 208 neurons) before declining to 96 at layer 21, whereas sadness signatures increased from 36 at layer 14 to 73 at layer 22. The corresponding English signature counts were 550, 114, 648, 635, 311, and 1,311. The smaller disgust set may partly reflect its lower sample size and the exclusion of small-effect associations by the corrected criterion.

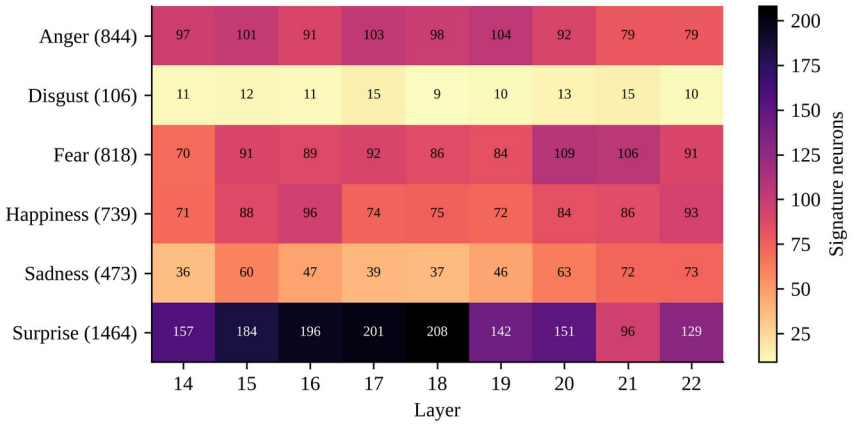


Figure 5. Corrected signature neurons per emotion at each layer (BH-FDR $q < 0.05$ within each layer and emotion, Cohen's h of at least 0.2), Chinese originals. Totals across layers are shown in parentheses.

4.2.2 *Split-half reliability of signature sets*

At layer 18, split-half reliability improved after correction for anger (0.733 versus 0.575), surprise (0.708 versus 0.563), fear (0.542 versus 0.453), and happiness (0.533 versus 0.504), while sadness remained ~ unchanged (0.397 versus 0.398) (Figure 6). Disgust had the lowest corrected reliability (0.219; standard deviation 0.068), potentially reflecting its smaller sample size of 1,290 texts, corresponding to 645 samples per split-half subset.

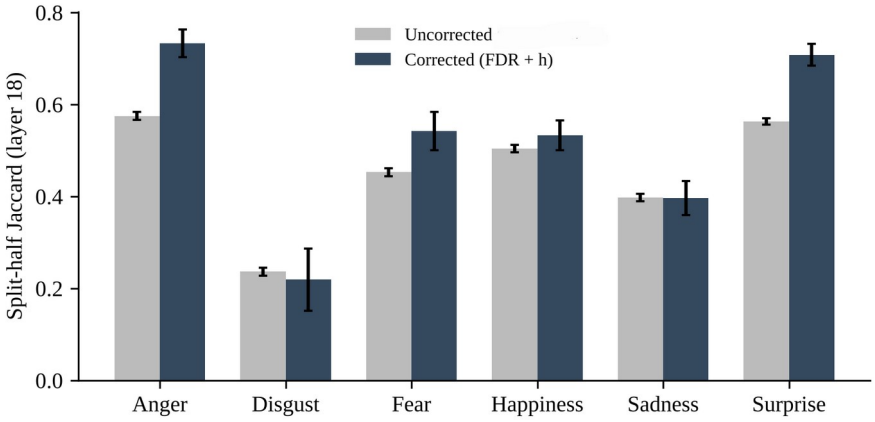


Figure 6. Split-half reliability of signature sets at layer 18 (100 random splits; mean and standard deviation), under the earlier uncorrected criterion and the corrected criterion.

4.2.3 *Cross-validated classification on binary FFN masks*

Five-fold cross-validated logistic regression on the re-extracted binary FFN activation masks achieved an accuracy of 0.526 (95% binomial confidence interval 0.518–0.534; N = 15,895), compared with 0.301 for majority-class prediction, 0.198 for proportional-chance prediction, and 0.167 for uniform guessing. Cohen's kappa was 0.399 and macro-averaged F1 was 0.464 (Figure 7). These results indicate modest emotion-discriminative information in the binary FFN activation patterns.

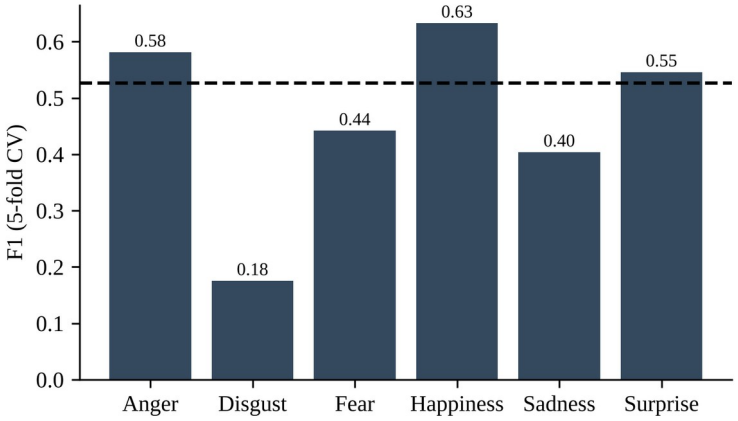


Figure 7. Per-emotion F1 scores of 5-fold cross-validated logistic regression on concatenated binary FFN masks (80,640 features, Chinese originals). The dashed line marks overall accuracy.

4.2.4 *Confusion patterns*

The confusion matrix (Figure 8) showed frequent classification of disgust (50%), sadness (26%), fear (23%), and surprise (22%) as anger. Happiness presented with the highest correct classification rate (67%). Because anger was the largest class, the analysis in Section 4.7 adjusted for marginal prediction frequencies when comparing confusion patterns with signature overlap.

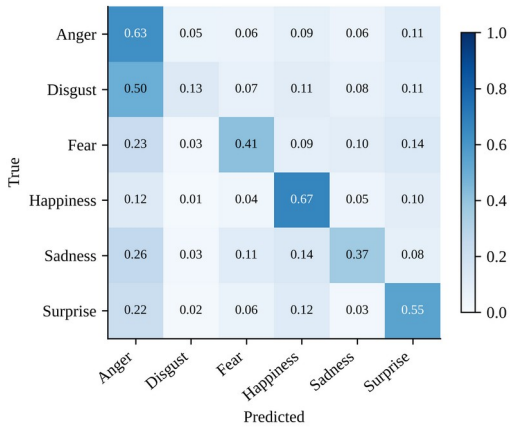


Figure 8. Row-normalised confusion matrix of the cross-validated classifier (Chinese originals).

4.3 Emotion-associated structure in the residual stream

Residual-stream activations from layers 14–22 showed limited separation of the six emotions. The first principal component explained 99.9% of the variance of the concatenated residual-stream vectors (99.8% in a re-extraction performed to verify this value). This figure was confirmed as computed, but it reflects the scale of a few units rather than emotion-related structure. Five hidden-state units (indices 286, 408, 520, 609 and 940) carried mean absolute activations between about 100 and 670 times that of the median unit in every examined layer, 45 of the 13,824 dimensions accounted for 92% of the total variance, and the first component was almost perfectly rank-correlated with the inverse of the token length of the text ($|\rho| = 0.999$). This is consistent with very large activations at a small number of tokens being divided by sequence length under mean pooling. After each dimension was standardized, the first component explained 55.8% of the variance and was again a length component ($|\rho| = 0.998$), and the second explained 1.7%. UMAP (Figure 9) and t-SNE projections showed substantial intermixing without visually distinguishable emotion clusters. The overall silhouette score was −0.090 (Figure 10; −0.080 in the re-extraction and −0.047 after standardization), indicating poor separation. Because the six classes are unequal in size, the purity expected when neighbors are unrelated to the label is the prevalence of each emotion in the corpus rather than 1/6; for anger, for example, it is 0.301. Measured against this baseline in the re-extraction, with each sample excluded from its own neighbors, 15-nearest-neighbor purity exceeded prevalence for every emotion (Figure 11): anger 0.444 against 0.301, disgust 0.116

against 0.081, fear 0.252 against 0.120, happiness 0.366 against 0.189, sadness 0.204 against 0.114, and surprise 0.389 against 0.195, that is, 1.4 to 2.1 times the baseline. The original analysis had counted each sample among its own 15 neighbors, which raised each value by 0.04 to 0.06 (to 0.177 for disgust and 0.485 for anger). The excess over prevalence indicates some local emotion-associated structure. Thus, although the residual stream did not exhibit separable emotion clusters under these measures, FFN gating activations yielded statistically reliable emotion signatures and above-baseline classification. This contrast concerns the detectability of emotion-associated structure in the examined activation spaces, not the absence of emotional information or computation in the residual stream.

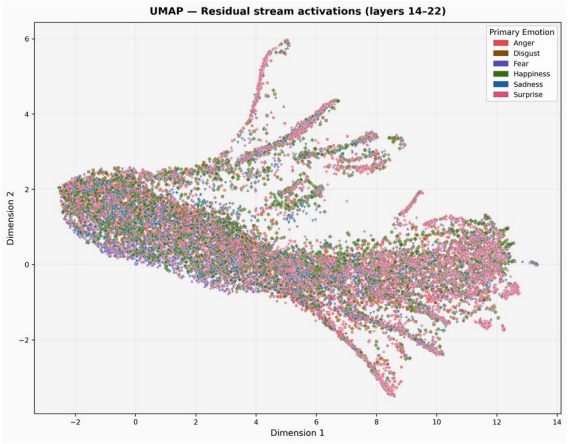


Figure 9. UMAP projection of concatenated residual stream activations (layers 14-22), coloured by primary emotion. The projection shows substantial intermixing of emotion labels, without clearly separable clusters under the visualization procedure employed.

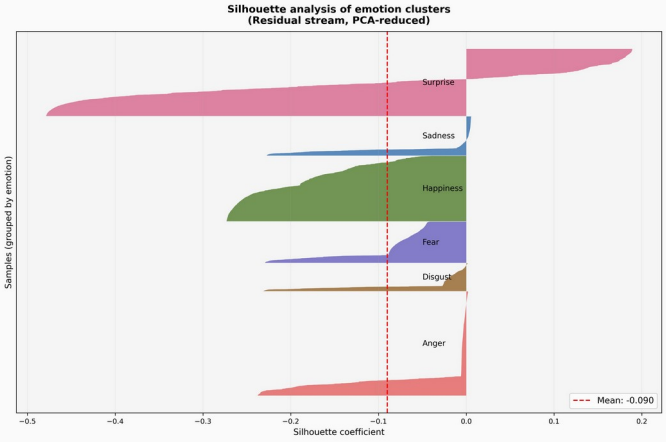


Figure 10. Silhouette analysis of emotion clusters in PCA-reduced residual stream space. The overall silhouette score of −0.090 indicates poor emotion-label separation under the specified analysis.

Figures 9 and 10 derive from the original extraction, which the verification re-extraction (CPU, bfloat16, identical pooling and truncation) reproduced closely; Figure 11 was regenerated from the re-extraction.

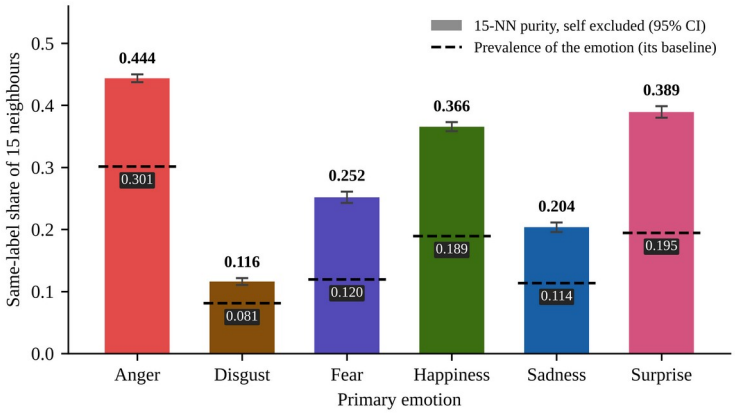


Figure 11. Fifteen-nearest-neighbor label purity per emotion in the 50-component PCA space of the residual stream (layers 14 to 22 concatenated; cosine distance; verification re-extraction), with each sample excluded from its own neighbors. Bars give the mean purity with 95% confidence intervals. The dashed line on each bar marks the prevalence of that emotion in the corpus (N = 15,895), which is its baseline, the purity expected if neighbors were unrelated to the label. Purity exceeds prevalence for every emotion, by a factor of 1.4 (disgust) to 2.1 (fear), indicating weak but consistent local structure.

4.4 Overlap structure of emotion signatures

4.4.1 *Pairwise Jaccard similarity*

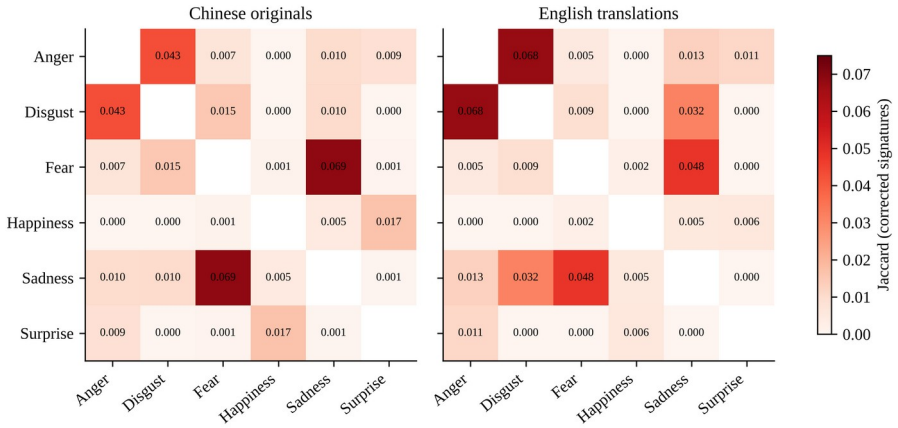


Figure 12. Pairwise Jaccard similarity of corrected signature sets aggregated over layers 14 to 22, for the Chinese originals (left) and the English translations (right).

The corrected Chinese signature-overlap matrix (Figure 12) showed the highest Jaccard similarities for fear–sadness (0.069), anger–disgust (0.043), and happiness–surprise (0.017), followed by disgust–fear (0.015) and disgust–sadness (0.010). Three pairs—anger–happiness, disgust–happiness, and disgust–surprise—shared no corrected signature neurons. Mean overlap was 11.9 times greater for same-valence than opposite-valence pairs, compared with 2.77 times under the earlier criterion. In the English activations, anger–disgust (0.068) and fear–sadness (0.048) remained the two highest-overlap pairs, whereas happiness–surprise overlap was lower (0.006) than disgust–sadness overlap (0.032).

4.4.2 *Layer-wise stability of overlap patterns*

The per-layer overlaps are less uniform under the corrected criterion than they had been. Fear and sadness range from 0.039 to 0.094 across the nine layers, anger and disgust from 0.011 to 0.074, and happiness and surprise from 0.005 to 0.032. Fear and sadness exceed happiness and surprise at every layer, and anger and disgust do so at eight of the nine. Per-layer values rest on small sets (for disgust, 9 to 15 neurons per layer), so the aggregate matrix is the more reliable summary.

4.4.3 *Exclusive versus shared neurons*

Figure 13 shows the balance between exclusive and shared corrected signature neurons. The proportion of exclusive neurons is 96% for surprise, 94% for happiness, 91% for anger, 87% for fear, 78% for sadness, 50% for disgust. The hierarchy reported earlier is preserved: surprise and happiness are the most self-contained, and disgust is by far the least autonomous, with half of its corrected signature neurons shared with other emotions, predominantly anger. All exclusivity ratios are higher than under the earlier criterion, because many of the neurons previously counted as shared were weakly associated with several emotions at once.

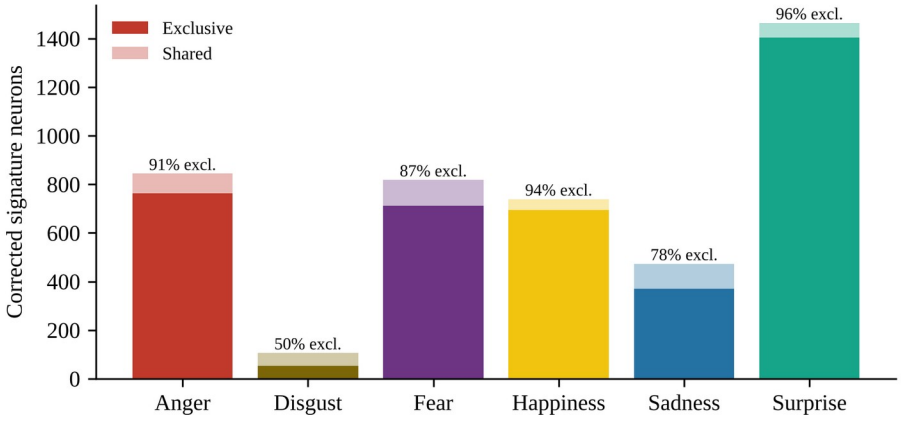


Figure 13. Exclusive (solid) and shared (translucent) corrected signature neurons per emotion (Chinese originals).

The largest pairwise intersections were fear-sadness [83 neurons], anger-disgust [39] and happiness-surprise [36], substantially below the counts obtained using the earlier uncorrected selection criterion.

4.4.4 *Hierarchical clustering*

Average-linkage clustering of the corrected Jaccard distances (Figure 14) merges fear and sadness first, then anger and disgust, then happiness and surprise, and joins the two negative pairs before either joins the positive pair. The cophenetic correlation is 0.990, which confirms that the tree represents the distance matrix faithfully.

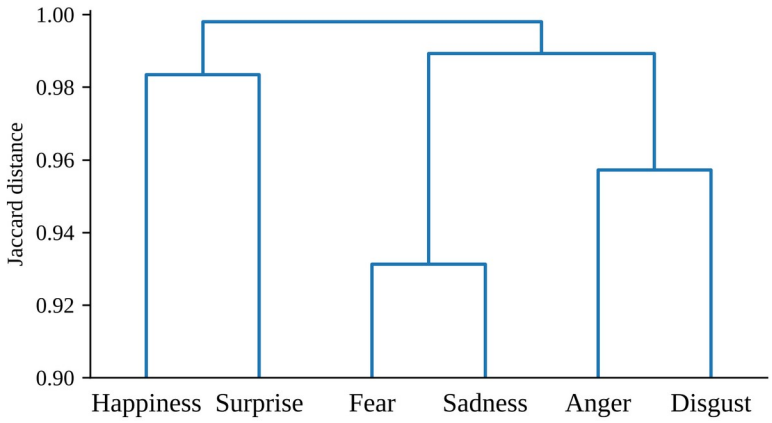


Figure 14. Hierarchical clustering of emotions by corrected signature overlap (average linkage, Jaccard distance, Chinese originals).

4.4.5 *Quantitative correspondence with the circumplex model*

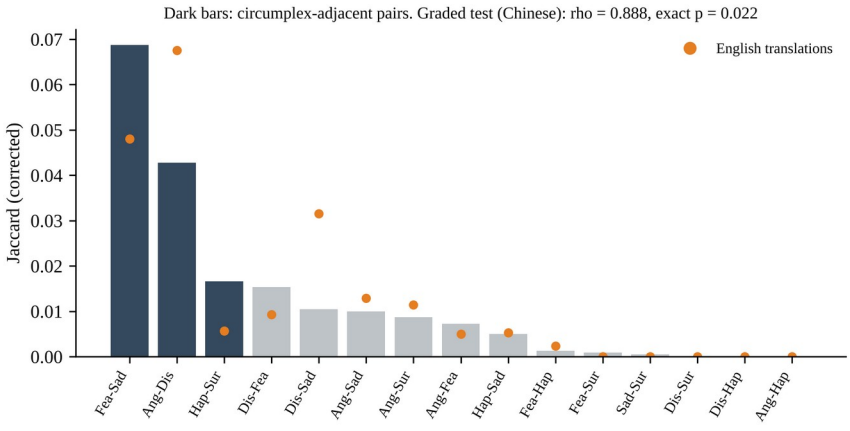


Figure 15. Corrected Jaccard similarity for all fifteen emotion pairs, ordered by magnitude (Chinese originals, bars) with the English values marked. The three pairs predicted to be adjacent under the circumplex model (dark bars) occupy the three highest Chinese positions. Exact permutation test over 720 relabelings: $\rho = 0.888$, $p = 0.022$.

Under the circumplex model three of the fifteen pairs are predicted to be adjacent: fear and sadness, anger and disgust, and happiness and surprise. With the corrected sets these three pairs again occupy the three highest positions (Figure 15); the probability of this under a random ordering is 1 in 455. Their mean overlap is 0.043 against 0.005 for the remaining twelve pairs. The Spearman correlation with the graded circumplex matrix is $\rho = 0.888$, and the exact permutation test over all 720 relabellings gives $p = 0.022$ (16 of 720). The binary quadrant model gives $\rho = 0.697$ (exact $p = 0.067$). In the English activations the graded correlation is $\rho = 0.773$ (exact $p = 0.022$) and the binary 0.550 ($p = 0.067$); the two negative pairs again rank first and second, but the third position is taken by disgust and sadness rather than happiness and surprise.

The exact permutation test is limited to 720 emotion-label relabelings, with a minimum attainable p-value of ~ 0.0014. The observed correspondence supports an association between neuronal overlap and the specified psychological emotion structure, but does not establish that the model computes emotions through those dimensions.

4.5 Sensitivity analysis

The analyses were repeated at activation thresholds of 1%, 3%, 5% and 10% (Figure 16). Consistent neuron counts scale almost linearly with the threshold and keep their relative ordering. The corrected signature totals also scale with the threshold, from 1,201 at 1% to 7,616 at 10%, with disgust smallest at every threshold. The corrected overlap matrix at each threshold correlates with that at 5% at $\rho = 0.920$ (1%), 0.964 (3%) and 0.996 (10%), and the graded circumplex correlation is 0.873, 0.815, 0.888 and 0.912 at the four thresholds, with exact $p = 0.022$ in every case.

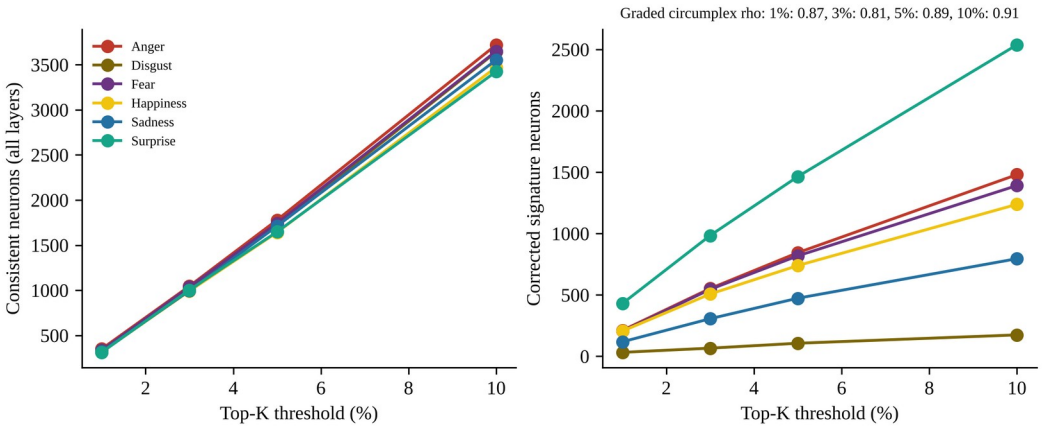


Figure 16. Left: consistent neuron counts as a function of the activation threshold. Right: corrected signature totals; the panel title gives the graded circumplex correlation at each threshold (Chinese originals).

4.6 Lexical similarity control

A lexical similarity matrix was computed from the English text of the corpus as described in Section 3.4.6. Lexical similarity carries some circumplex structure, but less than the corrected neuron overlap: its correlation with the graded circumplex model is ρ = 0.480 (exact p = 0.044), against 0.888 for the neuron overlap. The two matrices are moderately related (ρ = 0.473), and the neuron-level correspondence survives when lexical similarity is partialled out (partial ρ = 0.855, exact p = 0.022, 16 of 720 relabellings). The same holds for the English activations (partial ρ = 0.714, exact p = 0.022). On lexical similarity alone the third and fourth positions are occupied by two pairs that are not circumplex-adjacent and fear and sadness ranks only sixth, whereas the corrected gating overlap recovers all three predicted pairs in the top positions in the Chinese activations (Figure 17).

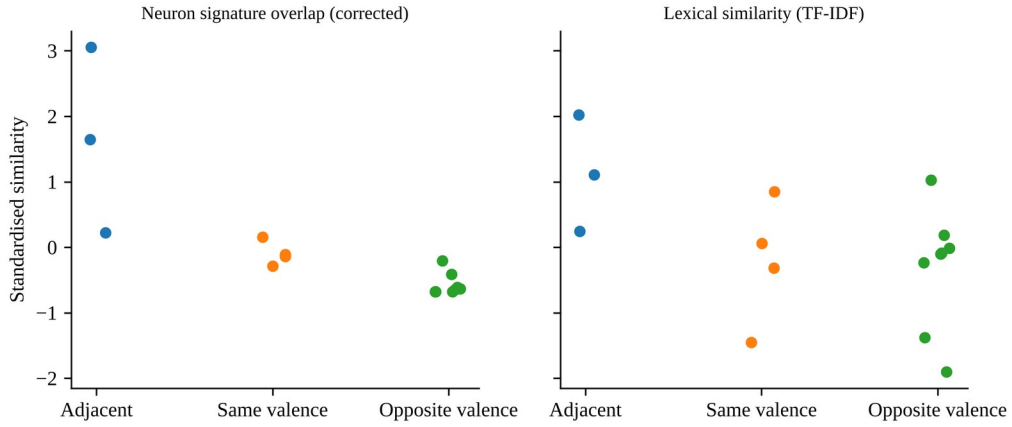


Figure 17. Standardised similarity between emotion classes grouped by circumplex relationship. Left: corrected neuron signature overlap. Right: lexical similarity of the same corpus. The neuron-level correspondence survives partialling out lexical similarity (partial ρ = 0.855, exact p = 0.022).

The circumplex correspondence therefore persists after controlling for the measured TF-IDF lexical similarity. However, this control does not exclude more complex lexical, syntactic, or semantic explanations.

4.7 Internal activation consistency

With corrected signature sets recomputed within each training fold, the mean signature-overlap matrix correlated with the marginal-frequency-corrected confusability of held-out classifier predictions (Spearman ρ = 0.789, exact permutation p = 0.0014; 1 of 720 relabelings; Figure 18). Because both the signatures and classifier features derive from FFN activation masks, this association demonstrates internal activation consistency rather than independent behavioral prediction. Behavioral correspondence is examined separately in Section 5.

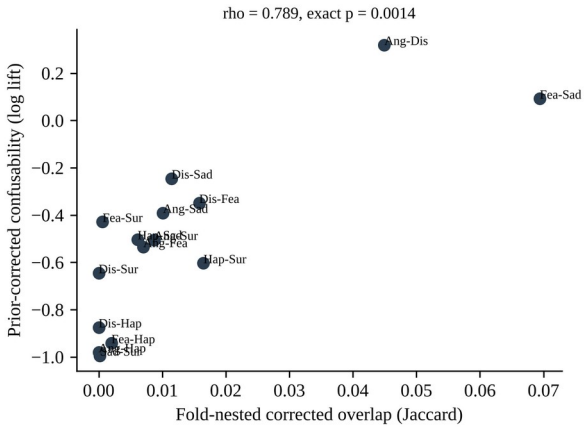


Figure 18. Fold-nested corrected signature overlap against marginal-frequency-corrected confusability of the activation-mask classifier for the fifteen emotion pairs (Chinese originals). Spearman ρ = 0.789, exact p = 0.0014. This is an internal activation-consistency result, not an independent behavioral validation.

4.8 Non-portability of neuron indices

The corrected signature neurons showed negligible index recurrence across layers: the 20 highest-ranked neurons for the same emotion shared an average of 0.042 indices between layers, compared with 0.040 expected by chance for randomly selected index sets of the same sizes (20 indices, or fewer where a layer's signature set was smaller, drawn from 8,960 possible indices), and 207 of 216 within-emotion layer pairs shared no indices. In contrast, within a single layer, the 20 most frequently activated neurons overlapped by an average of 17.5 across emotion pairs, indicating an emotion-general activation backbone rather than emotion-specific selectivity. These findings support the use of differential emotion signatures rather than raw activation rankings and demonstrate that neuron indices are meaningful only within their specific layer and model. The potentially transferable finding is the relative organization of emotion-associated overlap (Sections 4.4.5 and 4.6), rather than the numerical identity of individual neurons. The experiments in Section 5 used a separate reference map derived exclusively from the English-language derivation set of 11,126 texts, as specified in Section 3.6.2, not the full-corpus signature sets enumerated in Section 4.2.1. Generalizability of the overlap structure can be assessed using the protocol in Section 7.2.

4.9 Cross-language replication

The corrected emotion-signature overlap structure was strongly correlated between Chinese texts and their English translations (Spearman ρ = 0.920, exact permutation p = 0.0014; Figure 12). Individual emotion signatures also showed substantial cross-language recurrence, sharing 39–590 neurons depending on the emotion, with Jaccard similarities of 0.22–0.31, compared with 0.15 to 24 shared neurons expected from random sets of equivalent sizes drawn from 80,640 layer–neuron positions. The principal exception was the weaker happiness–surprise overlap in English. These findings demonstrate reproducibility of the emotion-associated

activation structure across the two language versions of the same texts, without establishing generalizability to independent corpora or models.

5. Results: reconfiguration experiments

5.1 Validity of the internal measure and deviations from preregistration

The preregistered internal measure scored a token by the mean of its signed, standardised activations over each emotion's signature neurons. The signature neurons, however, are defined by how often a neuron falls in the top 5% of activations by absolute value, so a neuron whose characteristic activations are strongly negative contributes to the signature when its signed value is low, and a signed average can cancel or invert the signal. This mismatch was found after Families 1 to 3 had been run and analysed with the preregistered measure. A validity check on the English held-out texts, which uses only baseline data, showed that the preregistered measure separated the six emotions at chance (six-way accuracy 0.165 against a majority rate of 0.302; within-pair AUC 0.44 to 0.51). Two measures consistent with the signature definition separated them well above chance: the fraction of an emotion's signature neurons that fall in the top 5% of the pooled activation, which applies exactly the binarisation used to define the signatures (six-way accuracy 0.425; within-pair AUC 0.68 to 0.83), and the mean standardised absolute activation over the signature neurons (0.429; 0.68 to 0.84). These accuracy and within-pair AUC results are distinct from the one-vs-rest AUC values plotted in Figure 19, which evaluate each emotion against the other five emotions combined.

The correction was written up as a deviation from the preregistration, and its document was hashed and time-stamped before any experimental result was computed with the corrected measures. The Deviation 1 hash record carries the UTC timestamp 5 October 2026 at 22:31:21, about two hours after the preregistration freeze at 20:36:36 (Section 3.6.1; Supplementary Section S7.1). At that time the results files of Families 1 to 3 were complete (last written at 21:00:34, 21:25:32 and 22:26:58 UTC), these families had been analysed with the preregistered measure, and the output of the held-out validity check had been written (22:29:46 UTC). Family 4 had not been completed: its results file was last written at 23:40:32 UTC, and its corrected-measure scores and all confirmatory result files were written after that. All Family 4 analyses, including those under the preregistered measure, were therefore run after the corrected measures had been fixed. These times are the file modification times preserved in the supplementary archive; separate run logs were not retained (Supplementary Section S11). A second deviation (Deviation 2) concerned only the software used to fit the linear mixed models (Section 3.6.5). Unlike Deviation 1, it was not written up, hashed or time-stamped when it was made; the archive file times place it at 21:15 UTC, after the Family 1 results were complete and before the Family 2 results were complete. The analysis script hashed at the preregistration freeze was modified after the freeze, and its released version therefore does not match the recorded hash; the documented modifications

are those of the two deviations, but the frozen version was not retained, so a line-by-line comparison cannot be supplied. All other files hashed at the freeze match their recorded hashes (Supplementary Section S11). The first of the two measures was fixed as the primary corrected measure because it is identical in construction to the signature definition, and the second as a robustness check. Because the pooled activations at every preregistered position had been saved during the experiments, the corrected scores were computed from those activations without rerunning the model, and every hypothesis, position, model, exclusion rule and the Holm family were left unchanged. All three sets of results are reported in Table 3, in the order in which the measures were adopted. The order of events matters for how the evidence should be weighed. The preregistered measure is the only measure that was fixed before any experimental data existed, so its results are the confirmatory results of this study in the strict sense. The preregistered-measure results for Families 1 to 3 had been seen before the correction was made. The protection against a measure chosen to suit the results is that the corrected measures were selected on held-out baseline data, by the criterion of consistency with the signature definition, and fixed and hashed before use; this makes the correction reasoned and documented, but it does not make the corrected analyses equivalent to an untouched prospective design. They are therefore described throughout as *post hoc* corrected analyses, and the conclusions that depend on them are qualified accordingly.



Figure 19. One-vs-rest AUC for the three internal measures on 4,769 held-out English texts without experimental prompts. The preregistered signed measure performed near chance, with AUC values both above and below 0.5, whereas the two post hoc corrected measures showed substantially higher discrimination. These one-vs-rest AUC values differ from the within-pair AUC ranges reported in Section 5.1.

Under the preregistered measure, 2 of the 14 hypotheses were supported after Holm correction (H4.2, H4.4). Under the primary *post hoc* corrected measure 10 were supported, and under the robustness measure 11. Because the preregistered measure did not measure what the hypotheses refer to, its largely null results are not evidence against the hypotheses, but neither are the corrected results confirmatory in the strict sense. The two corrected measures agreed

on 11 of the 14 decisions; they differed on H1.2, H2.2, H2.3, each supported by one corrected measure and not the other, and these results are treated below as not robust to the choice of measure. The results below refer to the primary *post hoc* corrected measure unless stated otherwise.

Table 3. Results of the 14 preregistered one-sided hypotheses under the three internal measures, in the order in which the measures were adopted. The preregistered measure is the confirmatory test as frozen; it failed the held-out validity check (Section 5.1). The two corrected measures were adopted afterwards through a documented deviation (Deviation 1), after the preregistered-measure results of Families 1 to 3 had been seen, and are therefore *post hoc* analyses rather than fully prospective confirmatory tests. Holm correction is applied across the 14 hypotheses within each measure. Prereg.: preregistered measure; Primary: primary corrected measure (mask fraction); Robust.: robustness corrected measure (absolute z); Held: supported after Holm correction. Mixed-model statistics are z = estimate / standard error with crossed random intercepts for item and paraphrase.

H	Prediction	Prereg.: Holm p	Held	Primary (*post hoc*): statistic	Holm p	Held	Robust. (*post hoc*): Holm p	Held
H1.1	Discrimination separates the pair more than recognition (stimulus tokens)	1.000	No	z = 4.27	7.7×10^{-5}	Yes	0.0007	Yes
H1.2	Paraphrase-level d' higher under discrimination (paraphrase-swap null)	1.000	No	sum of delta d' = 0.164	0.0003	Yes	0.557	No
H1.3	Internal separation predicts correct discrimination	1.000	No	z = 15.88	5.7×10^{-56}	Yes	2.3×10^{-58}	Yes
H1.4	Size of internal change predicts behavioural gain	1.000	No	z = 9.73	1.3×10^{-21}	Yes	3.5×10^{-42}	Yes
H2.1	Pair axis moves towards the evidenced emotion	1.000	No	z = 18.49	1.8×10^{-75}	Yes	2.2×10^{-182}	Yes
H2.2	Evidence shift exceeds shift from a neutral filler	1.000	No	W = 3856	0.750	No	8.0×10^{-11}	Yes
H2.3	Internal shift predicts behavioural shift	0.117	No	z = 0.58	0.750	No	0.0042	Yes
H3.1	Target emotion score exceeds the other five	1.000	No	z = 4.10	0.0001	Yes	0.0007	Yes
H3.2	Internal separability falls with reference overlap	1.000	No	ρ = -0.22	0.750	No	0.557	No
H3.3	Classifier separability falls with reference overlap	1.000	No	ρ = -0.27	0.750	No	0.557	No
H3.4	Target margin predicts classifier accuracy	1.000	No	z = 5.11	1.5×10^{-6}	Yes	2.1×10^{-7}	Yes
H4.2	Target score higher when expressed than suppressed	2.6×10^{-27}	Yes	z = 6.38	9.0×10^{-10}	Yes	9.7×10^{-15}	Yes
H4.3	Expression effect larger in output than at planning state	1.000	No	z = 3.49	0.0012	Yes	6.5×10^{-19}	Yes
H4.4	Internal target score predicts classifier probability	0.0496	Yes	z = 6.85	4.0×10^{-11}	Yes	1.5×10^{-8}	Yes

5.2 Family 1: recognition versus forced discrimination

Using the primary *post hoc* corrected measure, forced discrimination increased internal separation between the paired emotions over the stimulus tokens relative to broad emotion recognition (H1.1: estimate = 0.0037, z = 4.27, Holm-adjusted $p = 7.7 \times 10^{-5}$). The increase was positive across all four instruction paraphrases and replicated in odd- and even-numbered item halves (z = 3.17 and 2.90). The summed change in paraphrase-level d-prime was 0.164 (H1.2: exact $p = 5.2 \times 10^{-5}$), exceeding the variation obtained by exchanging condition labels among paraphrase cells. However, H1.2 did not survive correction under the alternative absolute-activation robustness measure (Holm p = 0.557), although the mixed-model result did. Pair-specific increases were detected for fear–sadness (z = 4.26) and anger–disgust (z = 2.05), but not happiness–surprise (z = 1.40). No difference was detected at the final prompt token (z = 0.04; Figure 20). The confidence intervals in Figure 20 describe the plotted item-level mean changes, whereas the pair-specific z-statistics are derived from mixed-effects models with item and paraphrase random effects. Consequently, whether a plotted interval includes zero need not agree with the significance of the corresponding mixed-model coefficient.

Internal separation was related to behaviour, in two ways that have to be distinguished. Within the discrimination condition, trials with greater internal separation towards the gold label were more often answered correctly (H1.3: z = 15.88), and an exploratory analysis showed that this held within each combination of pair and gold label (z = 9.72; robustness measure 8.31), so it is not a by-product of differences between emotions. Across items and paraphrases, the size of the internal change from recognition to discrimination predicted the size of the change in the behavioural margin (H1.4: z = 9.73; item-level Spearman ρ = 0.56, n = 480) (Figure 20). The same exploratory analysis showed, however, that this relation was carried mainly by a shift common to all items of one gold class. Under the discrimination instruction both the internal axis and the behavioural margin moved towards the label that the model favoured, disgust in the anger and disgust pair (mean internal change 0.063 for disgust items against -0.053 for anger items) and surprise in the happiness and surprise pair, so that items of the favoured class gained on both measures and items of the other class lost on both. Within each pair and gold-label cell the trial-level relation was weak (z = 2.12; robustness measure 0.97) and the item-level correlation was negative in five of the six cells. H1.4 met the originally specified statistical criterion under both post hoc corrected measures, but not under the preregistered measure. The association primarily reflects coordinated changes in internal activation and response preference, rather than evidence that items exhibiting greater representational reconfiguration were discriminated more accurately. The paraphrase-level d-prime of H1.2, which is unaffected by a shift common to both classes, is the cleaner evidence that the pair became more separable.

One result qualifies this picture. Mean forced-choice accuracy was not higher under discrimination than under recognition (for example 0.66 against 0.75 for fear and sadness).

An exploratory check showed why: when both labels were named in the instruction, the model showed strong label preferences, choosing disgust over anger on 84% of anger and disgust trials and surprise over happiness on 81%, irrespective of the gold label, and it slightly favoured whichever label was listed first (54% of choices). A response bias of this kind lowers forced-choice accuracy, and it is the same bias towards one label with which the internal state moved in the analysis of H1.4 above.

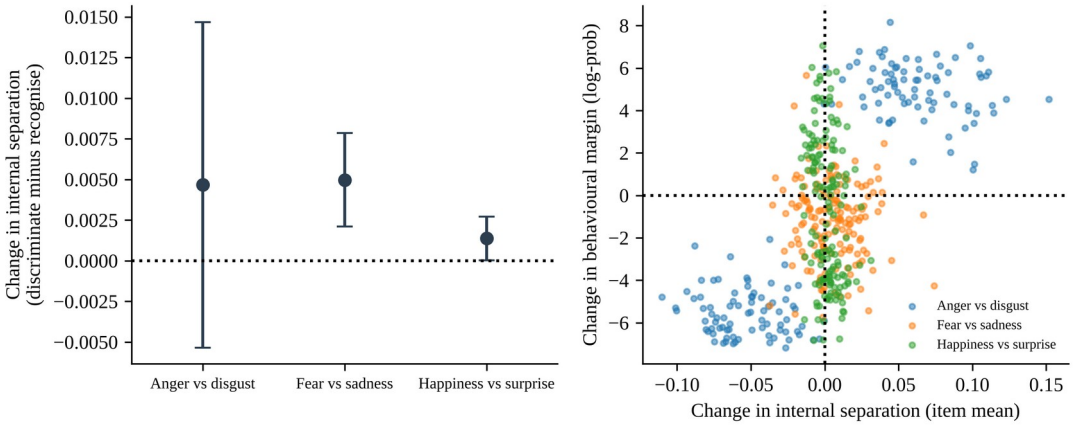


Figure 20. Family 1. Left: change in internal separation of the pair (discriminate minus recognise; item means with 95% confidence intervals). Right: item-level change in internal separation against change in the behavioural log-probability margin (H1.4).

5.3 Family 2: disambiguation after new evidence

Under the primary post hoc corrected measure, adding emotion-specific evidence was associated with a directional change in the internal pair axis (H2.1: $z = 18.49$), although the magnitude and symmetry of the response differed between the two evidence conditions and among emotion pairs (Figure 21). The internal shift was in the evidenced direction for 87% of item and paraphrase units, and the model's choice followed the evidence on 74% (anger and disgust), 90% (fear and sadness) and 96% (happiness and surprise) of evidence trials.

Under the primary *post hoc* corrected measure, two predicted effects were not supported. The evidence-associated shift (mean = 0.0063) did not exceed the absolute shift caused by a neutral filler sentence (mean = 0.0061; H2.2: Wilcoxon $p = 0.277$), and the magnitude of the internal shift did not predict the behavioral shift (H2.3: $z = 0.58$). The direction of the evidence-associated shift was appropriate, but its magnitude was not distinguishable from a generic sentence-induced perturbation. Under the alternative absolute-z robustness measure, both effects were supported (H2.2: Holm $p = 8.0 \times 10^{-11}$; H2.3: Holm $p = 0.0042$). Consequently, the directional response to evidence was robust, whereas evidence-specific magnitude and behavioral proportionality depended on the internal measure and remain unresolved.

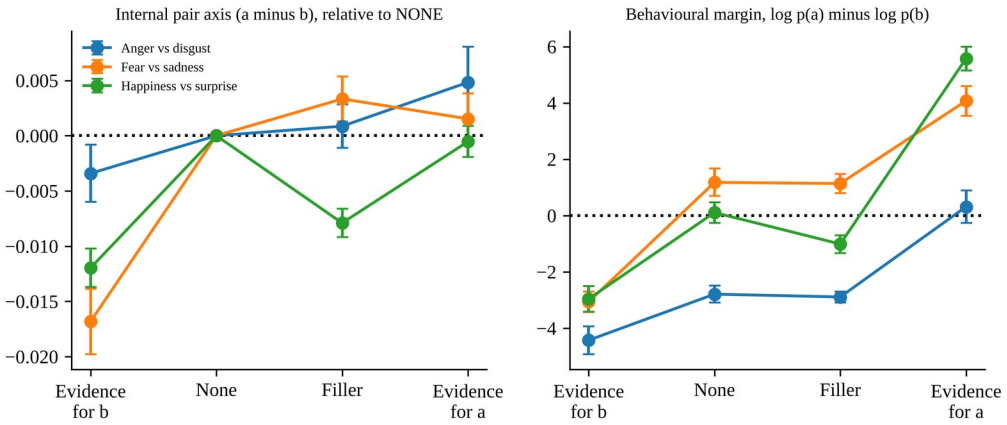


Figure 21. Family 2. Left: internal pair axis at the answer position relative to the base scenario, by condition (item means with 95% confidence intervals). Right: behavioural log-probability margin for the same conditions.

5.4 Family 3: targeted tone generation

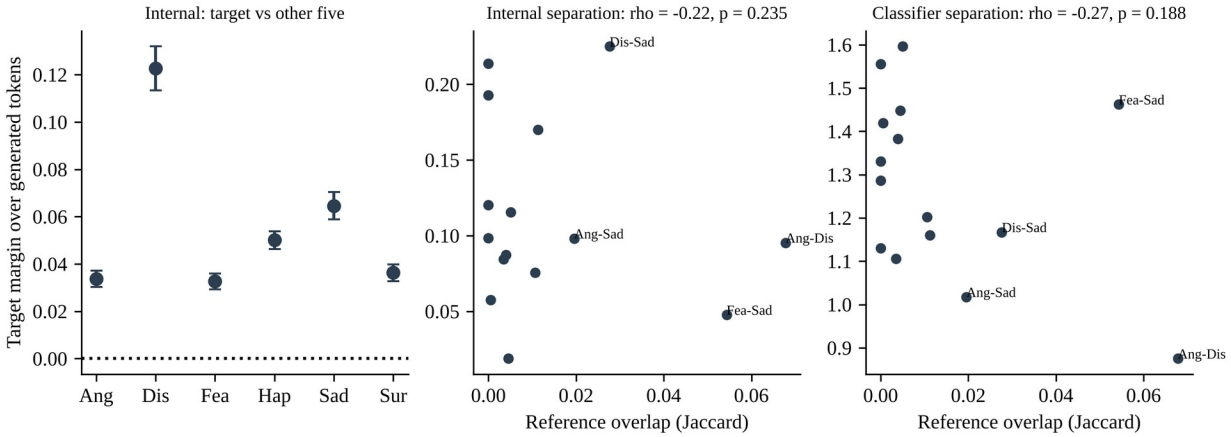


Figure 22. Family 3. Left: internal target margin over generated tokens by requested emotion (item means with 95% confidence intervals). Centre and right: internal and classifier separability of the fifteen emotion pairs against overlap in the reference map (H3.2 and H3.3).

When the model rewrote a neutral statement to express a requested emotion, the internal score of the target emotion over the generated tokens exceeded that of the other five (H3.1: z = 4.10), and the size of this target margin predicted whether the independent classifier recognised the intended emotion (H3.4: z = 5.11), also within each requested emotion (exploratory, z = 6.84; robustness measure 6.64). The classifier recognised the target in 72% of generations, from 46% for anger to 91% for sadness. The target margin was already present at the final prompt token before any text was generated (z = 3.35, secondary).

Greater reference-map overlap did not predict lower emotion-pair separability under the primary *post hoc* corrected internal measure (H3.2: Spearman ρ = −0.22, exact p = 0.235) or the preregistered behavioral classifier (H3.3: ρ = −0.27, p = 0.188; Figure 22). Although anger–disgust, the highest-overlap pair, was the least separable according to the classifier, the remaining pairs did not follow the predicted ordering. A secondary NRC lexicon analysis showed the predicted negative association (ρ = −0.75, exact p = 0.0028). Thus, the relationship between reference-map overlap and the emotional separability of generated text depended on the scoring instrument; the preregistered behavioral prediction was not supported by its designated classifier.

5.5 Family 4: expression versus suppression

The manipulation was effective: the classifier's probability for the target emotion was higher when the model was asked to express anger or fear than when it was asked to suppress it (z = 10.10; mean 0.57 against 0.16 for anger and 0.87 against 0.65 for fear), and the NRC lexicon agreed (z = 11.08). The internal target score over the generated text was correspondingly higher under expression (H4.2: z = 6.38), consistently across the four paraphrases, and across all three conditions it predicted the classifier's probability (H4.4: z = 6.85).

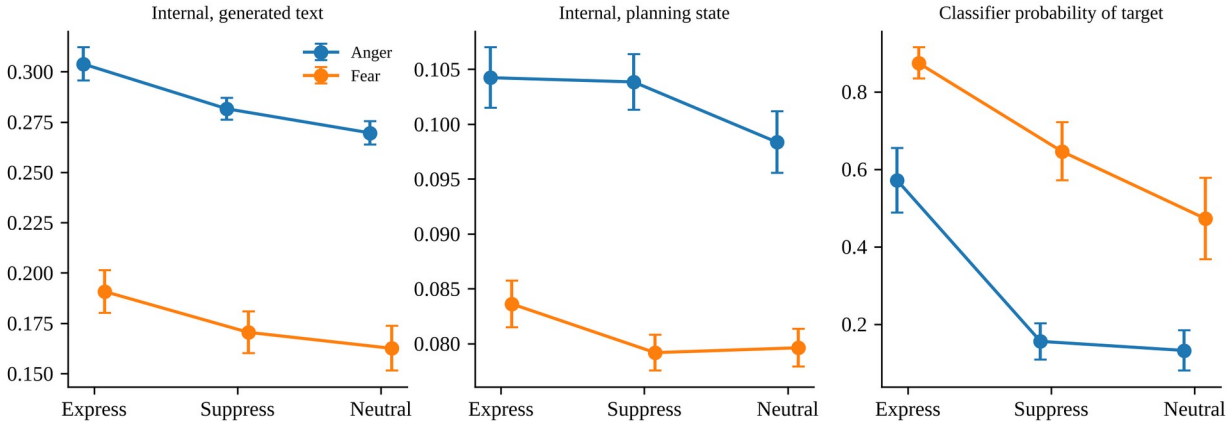


Figure 23. Family 4. Internal target score over the generated text (left) and at the planning state before generation (centre), and the classifier's probability of the target emotion (right), by condition (item means with 95% confidence intervals).

At the final prompt token before generation (the planning state), the primary *post hoc* corrected target score did not differ reliably between expression and suppression (z = 1.31; secondary analysis). The expression–suppression difference was greater over generated tokens than at the planning state (H4.3: z = 3.49; Figure 23). Suppressed responses retained greater target-emotion activation than neutral responses (z = 2.83; secondary), whereas their planning-state scores did not differ reliably (z = 1.09). These results are consistent with changes occurring during generation or through computations not captured by the measure. They

do not establish that the complete internal emotional representation remains unchanged during suppression, because the analysis samples only one pre-generation position and a binarized subset of activation dimensions.

5.6 Summary against the preregistered criteria

The experiments provide distinct levels of evidence. Statistically corrected emotion-associated signatures established representational structure, while post hoc corrected measures revealed task-dependent changes in their activation. Several internal–behavioral associations were detected, including relationships between activation measures and response preferences or generated emotional expression. However, these associations did not consistently establish the predicted behavioral consequences: greater internal separation did not improve forced-choice accuracy, evidence-specific effects depended on the measure, and reference-map overlap did not predict generated-text confusability under the preregistered classifier, although a secondary lexicon analysis supported the predicted relationship. The preregistered internal measure failed held-out validation, and only two of fourteen hypotheses were supported under that measure. Thus, the findings support identifiable emotion-associated representations and qualified evidence of their reconfiguration, but do not establish a general predictive mechanism or causal control over emotional behavior.

6. Discussion

6.1 Robustness, psychological correspondence, and limitations of the corrected emotion signatures

The correspondence with the circumplex model is particularly relevant because the language model was trained through next-token prediction rather than explicit supervision using emotion labels or emotional similarity judgments. The findings suggest that aspects of affective structure may be recoverable from statistical regularities in language. However, two interpretations remain possible: the model may have developed an organization of emotional information resembling relationships described by human affective models, or psychologically related emotions may occur in similar linguistic contexts, producing corresponding activation patterns without the model representing affective dimensions as such. The lexical analysis (Section 4.6) showed that the circumplex correspondence persisted after controlling for the bag-of-words similarity measure used in this study, although more complex semantic or contextual explanations remain possible.

The statistical correction also altered several conclusions from the earlier analysis, including the apparent depth dependence of surprise signatures and the robustness of happiness–surprise overlap across languages. Moreover, the cross-validated relationship between signature overlap and classifier confusion establishes internal activation consistency, not behavioral validation.

The corrected baseline therefore establishes statistically identifiable emotion-associated activation patterns whose overlap corresponds to the circumplex model and cannot be explained by the measured lexical similarity alone. It does not establish that the identified neurons represent emotions exclusively,

perform common emotional functions, or determine the model's behavioral responses.

6.2 Task-dependent reconfiguration and Its relationship to behavior

The experiments establish that emotion-associated activation patterns can change with task demands, but these changes do not consistently explain or predict behavioral performance. Detecting internal reconfiguration is therefore distinct from demonstrating its functional significance: greater separation or reorganization of emotion-associated signatures need not produce better emotional discrimination or more effective behavioral regulation. The relationship depends on the task, the internal measure employed, and the behavior being evaluated.

Measurement position and timing are particularly important. Changes detected at stimulus-token or answer positions may differ from those occurring during generation. Consequently, the absence of a reliably detected pre-generation activation difference between emotional expression and suppression does not establish that internal regulation is absent; relevant computations may occur later or outside the sampled, binarized activations. Comparisons with human expressive suppression (33) are suggestive but do not establish equivalent mechanisms. The experiments thus demonstrate task-dependent reconfiguration of the measured signatures without establishing a general predictive or causal relationship between that reconfiguration and emotional behavior.

6.3 Differences in emotion-associated structure between FFN Gates and the residual stream

Emotion-associated structure was more readily detectable in feed-forward gating activations than in the residual stream under the measurements employed. FFN activations yielded statistically reliable emotion signatures and above-baseline classification, whereas the residual-stream analysis produced a negative silhouette score and did not reveal separable emotion clusters. This difference is consistent with the interpretation of feed-forward layers as key-value memories (12), but does not establish that gating activations implement selective retrieval or causally determine emotional behavior. The findings also do not exclude emotional information in the residual stream or the possibility that FFN signatures reflect linguistic features correlated with emotion labels.

The layer-18 permutation tests identified a distinction in emotion-dependent activation frequency: anger, disgust, fear, and sadness showed counts exceeding their respective label-permutation null distributions, whereas happiness and surprise did not. This distinction should not be interpreted as evidence of fixed neuronal wiring or flexible computational routing. In particular, signature exclusivity and split-half reliability did not follow the same grouping: surprise exhibited high exclusivity and reliability, whereas disgust showed comparatively low values on both measures. These differences may reflect variation in corpus composition, annotation conventions, or linguistic context rather than distinct neural mechanisms. Thus, the permutation results establish an emotion-dependent statistical

pattern, not a mechanistic classification of emotions into fixed and flexible representations.

7. Limitations and future work

The principal limitation of this study is the absence of direct causal intervention on the identified emotion-associated neurons. The experiments manipulated task conditions rather than neuronal activations; no neuron-level ablation, activation steering, or other intervention was performed. Consequently, the observed associations and task-dependent changes do not establish that these neurons are necessary or sufficient for emotional behavior. Their activation may reflect emotion-related computations, correlated linguistic features, or downstream effects of computations elsewhere in the model. Accordingly, the identified signatures are statistical markers of emotion-associated activity, not demonstrated mechanisms of behavioral control.

7.1 Causal experiments that should follow

Three experiments would extend the present findings. First, selective ablation of emotion-associated signature neurons, through zeroing or mean substitution during generation, could test whether their disruption selectively alters the corresponding emotional behavior while preserving other emotional responses and general fluency, following the approaches of Vig et al. (17) and Arditi et al. (45). Second, activation steering could test whether amplifying these signatures predictably biases generated text toward the corresponding emotion, using methods developed by Rimsky et al. (43) and Turner et al. (44). Applying steering at the pre-generation planning state during suppression would additionally test whether emotional expression can be altered through intervention at that position. Third, cross-model comparisons involving Qwen, LLaMA, Gemma, Mistral, and Phi could examine whether similarities in emotion-associated neuronal organization correspond to similarities in classification accuracy, emotional consistency, controllability of tone, and responses to ambiguous prompts. Such comparisons would assess generalizability but remain correlational; ablation and steering would provide the direct tests of causal involvement absent from the present study.

7.2 A pre-specified protocol for testing generalisation

The regularity reported in Sections 4.4.5, 4.6 and 4.9 is a claim about the ordering of emotion pairs, not about neurons, and it is therefore testable on any model. This subsection specifies the test in advance so that a subsequent study can execute it without further decisions from the present authors, and so that the outcome cannot be adjusted after the fact.

The procedure comprises six steps. First, select a model exposing gated feed-forward layers and a corpus labeled with the same emotion taxonomy. Second, select a prespecified contiguous block of intermediate-to-later transformer layers (layers 14–22 of 28 in the present study), extract per-sample pooled activations, and binarize them at a top-K percentage threshold held fixed across emotions. Both the layer range and K should be selected before examining the results; sensitivity to K should be evaluated as in Section 4.5. Third, derive signature sets per emotion by one-sided proportion z-test against

all other emotions, retaining neurons with a Benjamini-Hochberg adjusted p below 0.05 within each layer and emotion and a Cohen's h of at least 0.2. Fourth, compute the pairwise Jaccard matrix over these sets. Fifth, correlate the matrix with the graded circumplex model matrix of Section 3.4.5 by Spearman rank correlation, with an exact p-value from all relabelings of the emotion set. Sixth, repeat the correlation partialling out a lexical similarity matrix from the same corpus. Any internal measure derived from the signatures should be validated on held-out data before being used to evaluate experimental hypotheses, as illustrated by the measure-validation results in Section 5.1.

The prediction is specific. If the regularity generalises, the three circumplex-adjacent pairs should occupy the highest positions of the ordering, the correlation with the graded model should exceed that with the binary quadrant model, and the correlation should survive the lexical control. The result that would falsify it is equally specific: an ordering in which the adjacent pairs are not elevated, or in which the correspondence disappears once lexical similarity is removed, would show that what is reported here is a property of Qwen2.5-1.5B-Instruct or of this corpus rather than a general feature of gated transformer language models. With six emotion categories, the exact permutation test permits 720 relabelings, giving a minimum attainable p-value of ~ 0.0014.

The protocol requires forward passes rather than gradient computation, model modification, or generation, making it less computationally demanding than the causal intervention experiments proposed in Section 7.1.

7.3 Further limitations

The findings are restricted to a single model, Qwen2.5-1.5B-Instruct, and nine of its 28 transformer layers. Although the baseline overlap structure was reproduced using English translations, this does not constitute replication in an independent corpus or model. Class imbalance, particularly the smaller disgust category, affects signature reliability. The reconfiguration experiments were conducted only in English, used greedy decoding, and relied on stimuli for Families 2 to 4 that were drafted by an AI system under the authors' direction rather than written by independent human authors or drawn from a corpus (Section 3.6.3). Behavioral evaluation also depended partly on an external emotion classifier whose agreement with the independent lexicon measure was incomplete.

A more consequential limitation concerns the internal activation measures. The preregistered measure failed its held-out validity check, necessitating *post hoc* corrections. Although the corrected measures revealed task-dependent effects, three hypothesis decisions differed between them. The original and corrected results are reported separately, but the corrected findings cannot be regarded as fully confirmatory. Furthermore, preregistration was documented through internal timestamps and cryptographic hashes rather than deposited in a public registry. These qualifications underscore the importance of validating internal measures independently before using them to test behavioral hypotheses.

Finally, the binary top-K activation criterion remains an operational simplification of

continuous neuronal activity, despite its demonstrated threshold robustness. Sparse autoencoder decomposition could provide a complementary characterization. The circumplex comparison involved only six emotions, limiting exact permutation testing to 720 relabelings. The revised CPU extraction closely reproduced the original activation statistics but not exactly. Figures 9 and 10 were not regenerated; the PCA, silhouette and nearest-neighbor purity results, including Figure 11, were recomputed from an independent re-extraction of the residual stream. Together, these limitations constrain the generalizability and measurement robustness of the findings without negating the reported statistical associations.

8. Conclusion

This study advances the interpretation of emotion-associated activity in language models by showing that statistically identifiable representations, their task-dependent reconfiguration, and their behavioral consequences are related but distinct phenomena. Corrected feed-forward activation signatures exhibited structured overlap corresponding to a psychological circumplex model, with substantial but incomplete cross-language reproducibility. The experimental findings revealed both meaningful internal–behavioral associations and important dissociations: internal activation changes tracked response preferences and aspects of generated emotional expression, yet increased internal separation did not improve forced-choice accuracy, and static signature overlap did not reliably predict generated-text confusability under the designated behavioral classifier. The latter relationship was nevertheless detected by a secondary lexicon measure, demonstrating sensitivity to the behavioral instrument. Crucially, the preregistered internal measure failed held-out validation, with only two of fourteen hypotheses supported; the evidence for task-dependent reconfiguration arose primarily from explicitly *post hoc* corrected measures. The manuscript's principal contribution is therefore not merely the identification of emotion-associated neurons, but the demonstration that representational organization and task responsiveness are insufficient, on their own, to establish behavioral predictability or causal control. By distinguishing these levels of explanation, the study identifies the limits of neuron-level emotion mapping and provides a framework for subsequent interventions testing how distributed, context-dependent computations generate emotional behavior.

Acknowledgements

The authors would like to express their sincere gratitude to Ms. Aashna Saraf and CreatED for providing the platform and opportunity to conduct this research.

Generative artificial intelligence tools were used during the preparation of this manuscript. In the present revision an AI system (Claude, Anthropic) was used, under the direction of the authors, to write and run the analysis and experiment code, to draft the experimental stimuli of Families 2 to 4 and the instruction paraphrases (Section 3.6.3; the Family 1 items are human-written texts from the dataset), to draft the preregistration and deviation documents, and to assist with drafting and

editing of prose. The study design, the research questions, the selection of model and dataset, the choice of task families and analysis plan, the interpretation of all results and the conclusions drawn are the work of the authors. All AI-assisted output, including every stimulus item, the preregistration and all code, was reviewed, verified and edited by the authors, who take full responsibility for the accuracy and integrity of the entire manuscript, including all statistical results and all references.

Data and code availability

The public data repository for all code used in this research is located at https://github.com/sankar-mechengg/CreatED-Rayan-Emotions-in-LLMs . In addition, all materials needed to reproduce the results are released as a supplementary archive under CC-BY-4.0: the extraction and analysis code for the corrected baseline in both languages; the nested classifier; the code that derives the reference map and splits the data; the four families' stimuli (the 480 held-out Family 1 items with their dataset row numbers, and the newly written items of Families 2 to 4); the experiment and analysis code, including the self-tests on synthetic data and the validation of the mixed-model implementation; the preregistration, Deviation 1 and their SHA-256 hash files; the model settings; the trial-level results of every experiment, including all generated text; and the signature enumerations and result tables. Cached activation arrays are not included because of their size and can be regenerated with the released code; they are available from the corresponding author on request.

9. References

1 Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word-emotion association lexicon. Computational Intelligence, 29(3), 436-465. doi:10.1111/j.1467-8640.2012.00460.x

2 Kim, Y. (2014). Convolutional neural networks for sentence classification. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 1746-1751). Association for Computational Linguistics. doi:10.3115/v1/D14-1181

3 Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (pp. 4171-4186). Association for Computational Linguistics. doi:10.18653/v1/N19-1423

4 Ekman, P. (1992). An argument for basic emotions. Cognition and Emotion, 6(3-4), 169-200. doi:10.1080/02699939208411068

5 Plutchik, R. (1980). A general psychoevolutionary theory of emotion. In R. Plutchik & H. Kellerman (Eds.), Theories of emotion (pp. 3-33). Academic Press.

6 Russell, J. A. (1980). A circumplex model of affect. Journal of Personality and Social Psychology, 39(6), 1161-1178. doi:10.1037/h0077714

7 Posner, J., Russell, J. A., & Peterson, B. S. (2005). The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. Development and Psychopathology, 17(3), 715-734. doi:10.1017/S0954579405050340

8 Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., & Ravi, S. (2020). GoEmotions: A dataset of fine-grained emotions. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 4040-4054). Association for Computational Linguistics. doi:10.18653/v1/2020.acl-main.372

9 Barrett, L. F. (2017). How emotions are made: The secret life of the brain. Houghton Mifflin Harcourt. ISBN 9780544133310.

10 Dai, D., Dong, L., Hao, Y., Sui, Z., Chang, B., & Wei, F. (2022). Knowledge neurons in pretrained transformers. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 8493-8502). Association for Computational Linguistics. doi:10.18653/v1/2022.acl-long.581

11 Meng, K., Bau, D., Andonian, A., & Belinkov, Y. (2022). Locating and editing factual associations in GPT. In Advances in Neural Information Processing Systems (Vol. 35, pp. 17359-17372). arXiv:2202.05262. doi:10.48550/arXiv.2202.05262

12 Geva, M., Schuster, R., Berant, J., & Levy, O. (2021). Transformer feed-forward layers are key-value memories. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (pp. 5484-5495). Association for Computational Linguistics. doi:10.18653/v1/2021.emnlp-main.446

13 Wang, K., Variengien, A., Conmy, A., Shlegeris, B., & Steinhardt, J. (2023). Interpretability in the wild: A circuit for indirect object identification in GPT-2 small. In Proceedings of the 11th International Conference on Learning Representations (ICLR 2023). arXiv:2211.00593. doi:10.48550/arXiv.2211.00593

14 Hanna, M., Liu, O., & Variengien, A. (2023). How does GPT-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model. In Advances in Neural Information Processing Systems (Vol. 36). arXiv:2305.00586. doi:10.48550/arXiv.2305.00586

15 Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., & Carter, S. (2020). Zoom in: An introduction to circuits. Distill, 5(3), e00024.001. doi:10.23915/distill.00024.001

16 Elhage, N., et al. (2021). A mathematical framework for transformer circuits. Transformer Circuits Thread. https://transformer-circuits.pub/2021/framework/index.html

17 Vig, J., Gehrmann, S., Belinkov, Y., Qian, S., Nevo, D., Singer, Y., & Shieber, S. (2020). Investigating gender bias in language models using causal mediation analysis. In Advances in Neural Information Processing Systems (Vol. 33, pp. 12388-12401). arXiv:2004.12265. doi:10.48550/arXiv.2004.12265

18 Cunningham, H., Ewart, A., Riggs, L., Huben, R., & Sharkey, L. (2024). Sparse autoencoders find highly interpretable features in language models. In Proceedings of the 12th International Conference on Learning Representations (ICLR 2024). arXiv:2309.08600. doi:10.48550/arXiv.2309.08600

19 Elhage, N., et al. (2022). Toy models of superposition. Transformer Circuits Thread. arXiv:2209.10652. doi:10.48550/arXiv.2209.10652

20 Bricken, T., et al. (2023). Towards monosemanticity: Decomposing language models with dictionary learning. Transformer Circuits Thread. https://transformer-circuits.pub/2023/monosemantic-features/index.html

21 Templeton, A., et al. (2024). Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. Transformer Circuits Thread. https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html

22 van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. Journal of Machine Learning Research, 9, 2579-2605. https://www.jmlr.org/papers/v9/vandermaaten08a.html

23 McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426. doi:10.48550/arXiv.1802.03426

24 Radford, A., Jozefowicz, R., & Sutskever, I. (2017). Learning to generate reviews and discovering sentiment. arXiv preprint arXiv:1704.01444. doi:10.48550/arXiv.1704.01444

25 Tigges, C., Hollinsworth, O. J., Geiger, A., & Nanda, N. (2023). Linear representations of sentiment in large language models. arXiv preprint arXiv:2310.15154. doi:10.48550/arXiv.2310.15154

26 Vaswani, A., et al. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (Vol. 30, pp. 5998-6008). arXiv:1706.03762. doi:10.48550/arXiv.1706.03762

27 Shazeer, N. (2020). GLU variants improve transformer. arXiv preprint arXiv:2002.05202. doi:10.48550/arXiv.2002.05202

28 Ramachandran, P., Zoph, B., & Le, Q. V. (2017). Searching for activation functions. arXiv preprint arXiv:1710.05941. doi:10.48550/arXiv.1710.05941

29 Touvron, H., et al. (2023). LLaMA: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971. doi:10.48550/arXiv.2302.13971

30 Qwen Team, et al. (2024). Qwen2.5 technical report. arXiv preprint arXiv:2412.15115. doi:10.48550/arXiv.2412.15115

31 Jiang, A. Q., et al. (2023). Mistral 7B. arXiv preprint arXiv:2310.06825. doi:10.48550/arXiv.2310.06825

32 Dauphin, Y. N., Fan, A., Auli, M., & Grangier, D. (2017). Language modeling with gated convolutional networks. In Proceedings of the 34th International Conference on Machine Learning (Vol. 70, pp. 933-941). PMLR. arXiv:1612.08083. doi:10.48550/arXiv.1612.08083

33 Gross, J. J. (1998). The emerging field of emotion regulation: An integrative review. Review of General Psychology, 2(3), 271-299. doi:10.1037/1089-2680.2.3.271

34 Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B (Methodological), 57(1), 289-300. doi:10.1111/j.2517-6161.1995.tb02031.x

35 Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Lawrence Erlbaum Associates.

36 Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with Python. In Proceedings of the 9th Python in Science Conference (SciPy 2010). doi:10.25080/Majora-92bf1922-011

37 Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. Proceedings of the National Academy of Sciences, 115(11), 2600-2606. doi:10.1073/pnas.1708274114

38 Hartmann, J. (2022). Emotion English DistilRoBERTa-base. Hugging Face model repository. https://huggingface.co/j-hartmann/emotion-english-distilroberta-base/

39 Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. Journal of Memory and Language, 59(4), 390-412. doi:10.1016/j.jml.2007.12.005

40 Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. Journal of Personality and Social Psychology, 103(1), 54-69. doi:10.1037/a0028347

41 Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. Journal of Statistical Software, 67(1), 1-48. doi:10.18637/jss.v067.i01

42 Holm, S. (1979). A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics, 6(2), 65-70.

43 Rimsky, N., Gabrieli, N., Schulz, J., Tong, M., Hubinger, E., & Turner, A. (2024). Steering Llama 2 via contrastive activation addition. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 15504-15522). Association for Computational Linguistics. doi:10.18653/v1/2024.acl-long.828

44 Turner, A. M., et al. (2023). Steering language models with activation engineering. arXiv preprint arXiv:2308.10248. doi:10.48550/arXiv.2308.10248

45 Arditi, A., et al. (2024). Refusal in language models is mediated by a single direction. In Advances in Neural Information Processing Systems (Vol. 37). arXiv:2406.11717. doi:10.48550/arXiv.2406.11717

46 Geiger, A., Lu, H., Icard, T., & Potts, C. (2021). Causal abstractions of neural networks. In Advances in Neural Information Processing Systems (Vol. 34, pp. 9574-9586). arXiv:2106.02997. doi:10.48550/arXiv.2106.02997

47 Belinkov, Y. (2022). Probing classifiers: Promises, shortcomings, and advances. Computational Linguistics, 48(1), 207-219. doi:10.1162/coli_a_00422

48 Stearns, F. W. (2010). One hundred years of pleiotropy: A retrospective. Genetics, 186(3), 767-773. doi:10.1534/genetics.110.122549

49 Edelman, G. M., & Gally, J. A. (2001). Degeneracy and complexity in biological systems. Proceedings of the National Academy of Sciences, 98(24), 13763-13768. doi:10.1073/pnas.231499798

50 Marder, E., & Goaillard, J.-M. (2006). Variability, compensation and homeostasis in neuron and network function. Nature Reviews Neuroscience, 7(7), 563-574. doi:10.1038/nrn1949