Decodable but Not Reliably Causally Operative:
A Controlled Scale Study of Emotion-Related Representations in Sub-2B Language Models

AI-Generated Experimental Research Report

Research system: OpenAI GPT-5.6 Sol
(high-reasoning configuration, operating through Codex)

Human initiator and artifact custodian: Josiah Wilson

July 25, 2026

Verification status. This report was designed, implemented, executed, analyzed, and drafted by an AI system. It has not been peer reviewed or independently validated by a domain expert. Josiah Wilson initiated the research question, authorized the computational work and public release, and preserves the artifacts; he does not claim technical authorship or independent verification of the mathematical and statistical content.

Abstract

Recent work has reported localized emotion representations and causally effective "emotion circuits" in large language models (LLMs), but linear decodability alone does not establish that a representation controls model behavior. We test whether emotion-related information becomes causally operative between 0.5B and 1.5B parameters in the Qwen2.5-Instruct family. Using 720 balanced, lexical-cue-filtered crowd-enVENT narratives for training and 311 temporally separated narratives for evaluation, layerwise linear probes reached held-out balanced accuracies of 0.601 (95% bootstrap interval 0.545–0.657) at 0.5B and 0.701 (0.649–0.751) at 1.5B, versus chance 1/6. Broader lexical filtering preserved the result. We then estimated matched, context-centered emotion directions from an independent controlled scenario dataset and intervened on five residual-stream layers during greedy generation. Across 36 untouched test scenarios, mean target-emotion probability changed by only +0.0133 at 0.5B and +0.0015 at 1.5B. No emotion satisfied prespecified sufficiency criteria against unsteered, norm-matched random, and label-shuffled controls after Holm correction. Projection removal likewise failed the necessity criterion, and selected interventions failed the prespecified fluency constraint. Thus, model scale improved decodability without improving causal steerability in this range. The results support a conservative distinction between emotion-related information and an emotion circuit; they do not imply subjective emotion or exclude nonlinear, distributed, or larger-model mechanisms.

1 Introduction

Statements that an LLM contains an emotion "circuit" can denote importantly different findings. A classifier may decode an emotion label from hidden states; a direction in activation space may

steer generated affect; or a sparse set of components may be both necessary and sufficient for an emotion-related computation. Only the latter findings warrant a mechanistic claim, and none implies phenomenal feeling.

Recent studies have localized emotion inference in autoregressive LLMs [7], identified shared late-layer emotion representations [5], and reported neuron–head circuits that control emotional expression in Llama-3.2-3B-Instruct [9]. These results motivate a scale question that has not been isolated: when, if at all, does a readily decodable emotion representation become a causally reliable mechanism?

The distinction matters because probing can establish accessibility to an auxiliary decoder without establishing use by the model [2, 4]. We therefore separate representation evidence from intervention evidence in a prospectively specified, within-family comparison of Qwen2.5-0.5B-Instruct and Qwen2.5-1.5B-Instruct [6]. The study combines temporally separated narrative data, fixed linear probes, matched-context direction estimation, held-out interventions, two control families, projection removal, an independent output evaluator, and a utility constraint.

Our central result is a dissociation: held-out emotion decodability increased by approximately ten percentage points with scale, whereas the mean causal steering effect decreased. Neither model met the prespecified definition of a causal emotion-related mechanism or candidate emotion circuit.

Contributions. This study (i) specifies an explicit evidential ladder from decodability to candidate circuit status; (ii) provides a reproducible sub-2B scale comparison with random and label-shuffled interventions; and (iii) reports a negative causal result despite strong positive probing results, illustrating why these forms of evidence should not be conflated.

2 Related Work

Emotion representations and mechanisms. Tak et al. found functionally localized representations during emotion inference across model families and used appraisal interventions to steer predictions [7]. Maheswaran and Desarkar reported common late-layer emotion representations whose directions transfer across datasets [5]. Most directly, Wang et al. introduced the Scenario–Event with Valence (SEV) dataset, estimated context-agnostic directions, decomposed neuron and attention-head contributions, and reported high-accuracy circuit steering in a 3B model [9]. We use their released SEV split and reproduce the matched-centering principle while testing smaller models and stronger held-out controls.

Causal probing. Representation engineering treats population-level directions as objects for monitoring and intervention [10]. However, probe capacity and control-task performance complicate representational claims [4], and successful decoding need not predict behavioral importance [2]. Causal probing itself has a completeness–selectivity trade-off [1]. These concerns motivate our separate tests of decodability, sufficiency, necessity, specificity, and fluency.

3 Research Questions and Evidential Criteria

We ask whether increasing model size from 0.5B to 1.5B parameters produces a transition from linearly decodable emotion information to a causally operative mechanism. Six categories are shared across all analyses: anger, disgust, fear, happiness, sadness, and surprise.

The prospective analysis plan defined:

1. **H1 (decodability):** a non-embedding layer has a lower 95% confidence bound above balanced chance 1/6;
2. **H2 (localization):** the selected layer exceeds the embedding baseline by at least 0.05;
3. **H3 (sufficiency):** target-direction addition increases the target evaluator probability relative to baseline, random directions, and label-shuffled directions;
4. **H4 (necessity):** removing the target-direction projection reduces target probability more than a random projection;
5. **H5 (specificity and utility):** target change exceeds non-target change, with neutral-reference perplexity degradation at most 15%; and
6. **H6 (scale):** the prespecified causal summary is monotonically nondecreasing with scale.

We reserve *emotion circuit* for a sparse mechanism passing necessity, sufficiency, specificity, out-of-domain generalization, and random-component controls. The present tests evaluate the prerequisites for component tracing; a failed global mechanism is not relabeled as a circuit through post-hoc component selection.

4 Methods

4.1 Models and execution environment

We evaluated the instruction-tuned Qwen2.5 checkpoints at 0.5B and 1.5B parameters [6]. Model revisions were pinned by artifact manifests. Inference and activation storage used float32. Experiments ran with PyTorch 2.7.1 and Transformers 4.53.3 on an Apple M4 Mac mini with 16 GB unified memory using the Metal backend. The global random seed was 20260725.

Llama-3.2-1B-Instruct was retained as an exact reproduction target, but its weights required authenticated gated access unavailable in the execution environment. It was therefore not silently substituted into the confirmatory analysis.

4.2 Recognition data and privacy-minimized preparation

The recognition experiment used the generation portion of crowd-enVENT, a corpus of human event descriptions paired with emotions and appraisals [8]. Only round identifier, text identifier, emotion, and generated narrative were read. Participant identifiers, demographics, and timestamps were excluded from all study artifacts.

Rounds 2–5 formed the training pool and round 6 the held-out temporal split. Rounds 7–9 were excluded before inference because several confirmatory classes were absent or extremely sparse. We removed missing narratives, duplicates, narratives shorter than four words, and examples containing the class name or a direct morphological form. Training activations used a deterministic balanced sample of 120 examples per class ($n = 720$). The held-out split retained all 311 eligible narratives. Happiness was mapped from the corpus label *joy*.

4.3 Layerwise representation probes

Each narrative was inserted into a fixed chat prompt requesting one of the six labels. We recorded the residual stream at the final non-padding input token for the embedding output and every transformer block output. At each layer, a pipeline standardized dimensions using training statistics

and fit an L2-regularized multinomial logistic regression ($C = 1$, class-balanced weights, maximum 4000 iterations). The layer was selected by five-fold stratified cross-validation on training data only, then evaluated once on round 6. The primary statistic was balanced accuracy. Intervals used 2,000 class-stratified bootstrap samples.

The lexical robustness analysis removed a broader prespecified set of direct synonyms (e.g., *furious*, *terrified*, *delighted*, *miserable*, and *astonished*), rebalanced training to 111 examples per class, and retained 305 validation examples. A five-repeat shuffled-training-label control was evaluated at the primary selected layer.

4.4 Matched emotion directions

Direction estimation used only the released SEV training split [9]. We deterministically sampled 24 skeletons and used each neutral event. For every skeleton g and emotion e, the model generated a response under an explicit emotion-style instruction. We then extracted the final-token residual activation $\mathbf{h}_{g,e}^{(\ell)}$ from the complete prompt–response transcript.

To remove context shared within a skeleton, activations were centered:

$$\tilde{\mathbf{h}}_{g,e}^{(\ell)} = \mathbf{h}_{g,e}^{(\ell)} - \frac{1}{6}\sum_{e'} \mathbf{h}_{g,e'}^{(\ell)}. \tag{1}$$

The cross-context centroid for each emotion was again centered across emotions and normalized:

$$\mathbf{v}_e^{(\ell)} = \frac{\mathbb{E}_g[\tilde{\mathbf{h}}_{g,e}^{(\ell)}] - \mathbb{E}_{e',g}[\tilde{\mathbf{h}}_{g,e'}^{(\ell)}]}{\left\| \mathbb{E}_g[\tilde{\mathbf{h}}_{g,e}^{(\ell)}] - \mathbb{E}_{e',g}[\tilde{\mathbf{h}}_{g,e'}^{(\ell)}] \right\|_2}. \tag{2}$$

Five within-skeleton label permutations produced shuffled directions, and five seeded Gaussian draws produced unit-norm random directions.

4.5 Causal sufficiency intervention

The intervention window comprised five hidden-state outputs centered on the training-CV-selected probe layer: indices 15–19 for 0.5B and 24–28 for 1.5B. At each greedy decoding step, the hook changed the final-token residual $\mathbf{h}$ according to

$$\mathbf{h}' = \mathbf{h} + \alpha \ \mathrm{RMS}(\mathbf{h})\, \mathbf{v}_e. \tag{3}$$

Strengths $\alpha \in \{0.25, 0.5, 1.0, 1.5, 2.0\}$ were evaluated on 12 SEV training skeletons disjoint from direction estimation. If the maximum development gain was positive, the smallest strength attaining 90% of it was selected. The prespecified rule was undefined when every gain was negative; before held-out 1.5B generation, it was amended to select the least negative strength, with ties favoring smaller α.

The untouched SEV test split supplied 36 deterministic neutral-event skeletons. For each target emotion, we generated an unsteered response, a target-steered response, five random-direction responses, and five label-shuffled-direction responses, all with greedy decoding and at most 24 new tokens. This yielded 2,412 scored rows per model with no empty responses.

Table 1: Held-out representation-probe results. Strict filtering removes a broader synonym list. Chance is 0.167.

Model	Selected layer	Primary BA	95% interval	Strict-filter BA
Qwen2.5-0.5B	17/24	0.601	[0.545, 0.657]	0.566
Qwen2.5-1.5B	27/28	0.701	[0.649, 0.751]	0.684

4.6 Outcome measurement and statistics

Generated text was scored by the independently trained Hartmann DistilRoBERTa emotion classifier, which predicts Ekman's six categories plus neutral [3]. Its *joy* output was mapped to happiness; probabilities were not renormalized after omitting neutral. We recorded target-emotion probability and all five non-target probabilities.

For each emotion, paired differences compared target steering with baseline, the mean of five random controls, and the mean of five shuffled controls. Intervals used 2,000 bootstrap resamples of scenario skeletons. One-sided Wilcoxon signed-rank tests were Holm-corrected jointly across the 18 emotion–comparison tests. H3 required a positive interval and adjusted $p < 0.05$ for every comparison.

4.7 Necessity and utility

Necessity used 18 held-out skeletons with explicit emotion instructions. At the same layer window, target projection removal applied

$$\mathbf{h}' = \mathbf{h} - (\mathbf{h}^\top \mathbf{v}_e)\mathbf{v}_e. \tag{4}$$

Outputs were compared with unmodified explicit generation and the mean of five random-projection controls using the same bootstrap and correction procedure.

For a utility stress test, we computed teacher-forced perplexity on 60 held-out neutral SEV prompts, adding the selected direction at all token positions. H5 required every emotion's mean per-example perplexity ratio to be at most 1.15.

5 Results

5.1 Emotion information was robustly decodable

Both models passed H1 and H2 (Table 1; Figure 1). The selected 0.5B layer achieved 0.601 balanced accuracy (95% interval 0.545–0.657), and the selected 1.5B layer achieved 0.701 (0.649–0.751). Embedding-layer performance was exactly chance in both models because the fixed final prompt token was identical across examples. Selected-layer improvements over this baseline were 0.435 and 0.534.

Broad lexical filtering retained lower interval bounds of 0.507 at 0.5B and 0.630 at 1.5B. Mean shuffled-label validation accuracy was 0.175 and 0.166, respectively. Thus, the result was not explained by direct emotion words or the fixed probe's capacity alone.

5.2 Explicit expression was imperfect in both small models

The independent evaluator's balanced accuracy on the 144 explicit outputs used for direction estimation was 0.438 for 0.5B and 0.514 for 1.5B. On held-out explicit baseline generations from the necessity experiment, it was 0.370 and 0.426. Both exceeded chance but revealed a strong

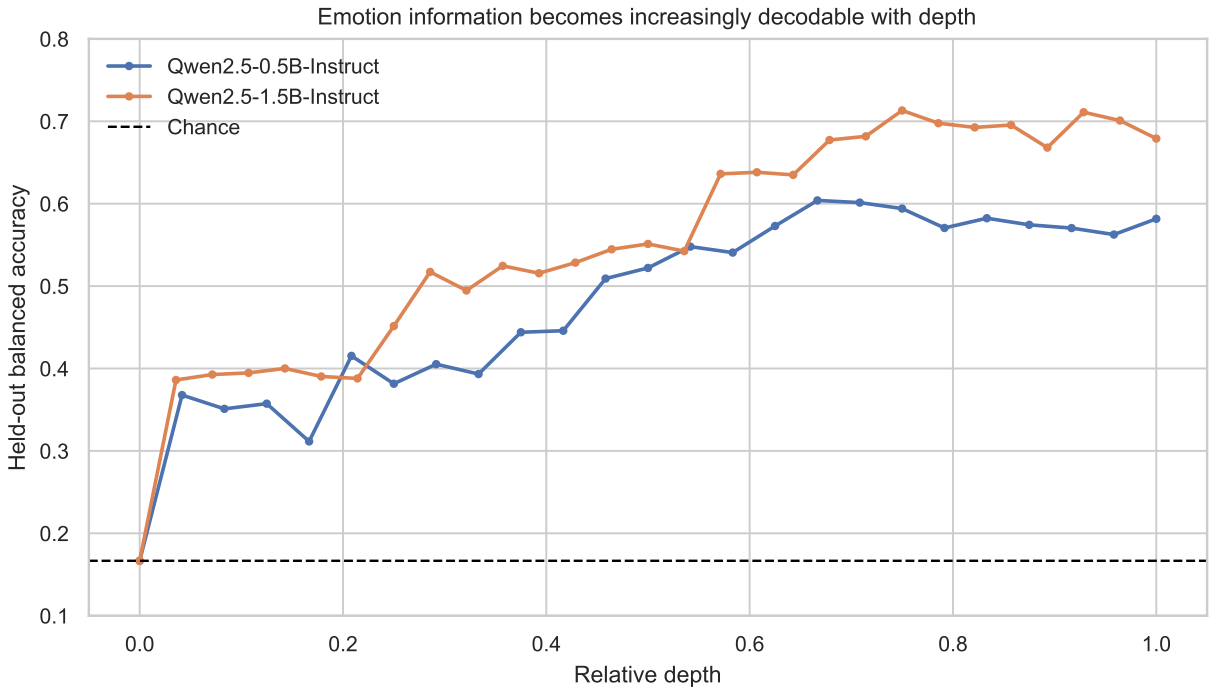

Figure 1: Held-out balanced accuracy by relative depth. Layer selection used training cross-validation, not the displayed validation values.

tendency to produce positive-sounding text, particularly for anger and fear prompts. This validation is important: direction estimates came from distinguishable but imperfectly elicited emotional expression.

5.3 Direction addition failed causal sufficiency

The 0.5B development curve was positive and selected $\alpha = 2.0$; the 1.5B curve was negative at every tested strength and selected the least negative value, $\alpha = 1.5$. On held-out scenarios, mean target-probability change relative to baseline was 0.0133 for 0.5B and 0.00155 for 1.5B. Table 2 reports per-emotion effects. No emotion in either model passed all three corrected comparisons; H3 failed.

Mean target-minus-random effects were 0.0284 and 0.0070; mean target-minus-shuffled effects were 0.0086 and 0.0105. The apparently larger 0.5B happiness effect had a wide interval and failed the full corrected criterion. Figure 2 displays the baseline contrasts.

5.4 Projection removal failed causal necessity

For 0.5B, the mean baseline-minus-target-projection contrast across emotions was 0.0368, but effects were heterogeneous: removing the sadness direction increased, rather than decreased, the target score, while surprise showed the largest predicted reduction. Only the surprise comparison against random projection survived correction; its baseline comparison did not. For 1.5B, the mean contrast was -0.1043, indicating that target projection removal often increased the evaluator's target probability. No emotion passed both necessity comparisons, so H4 failed in both models.

Table 2: Held-out target-probability change relative to unsteered baseline. Intervals are scenario-bootstrap 95% intervals.

Emotion	Qwen2.5-0.5B	Qwen2.5-1.5B
Anger	0.0066 [0.0002, 0.0176]	-0.0010 [-0.0035, 0.0016]
Disgust	0.0038 [-0.0016, 0.0101]	-0.0029 [-0.0155, 0.0047]
Fear	0.0063 [-0.0012, 0.0188]	0.0117 [-0.0082, 0.0431]
Happiness	0.0766 [-0.0201, 0.1809]	-0.0010 [-0.0662, 0.0650]
Sadness	0.0013 [-0.0006, 0.0033]	-0.0017 [-0.0051, 0.0006]
Surprise	-0.0152 [-0.0601, 0.0112]	0.0041 [-0.0095, 0.0175]
Mean	0.0133	0.0015

Table 3: Disposition of prospectively specified hypotheses.

Hypothesis	0.5B	1.5B
H1 Decodability	Supported	Supported
H2 Layer localization	Supported	Supported
H3 Causal sufficiency	Not supported	Not supported
H4 Causal necessity	Not supported	Not supported
H5 Specificity/utility	Not supported	Not supported
H6 Monotonic causal scale effect	Not supported	

5.5 Specificity, utility, and scale criteria failed

Specificity contrasts were unstable across emotions, and the utility constraint failed. Specificity intervals excluded zero for three of six 0.5B directions (disgust, fear, and sadness) and zero of six 1.5B directions; in the former cases, the contrast was driven primarily by reduced non-target scores rather than a reliable increase in the target score. Mean perplexity ratios were 1.613 for 0.5B and 1.083 for 1.5B, but H5 required the 1.15 limit for every emotion. Five of six 0.5B directions and three of six 1.5B directions exceeded the limit. Finally, the causal summary decreased from 0.0133 to 0.00155 while probe accuracy increased from 0.601 to 0.701. H6 therefore failed.

6 Discussion

The scale comparison shows that stronger, later-layer decodability is not equivalent to stronger causal control. The 1.5B model encoded the six-way emotion label substantially better than the 0.5B model, including after broad cue filtering, yet its matched residual directions did not provide stronger held-out steering. This is consistent with the general caution that a probe may recover information that downstream computation does not use in the tested linear form [2].

The negative result should be interpreted at the level of the intervention. It rules against a reliable, context-agnostic, linear residual-direction mechanism at the selected layers and strengths in these two checkpoints. It does not establish that the models lack all causally relevant emotion computation. Candidate mechanisms may be nonlinear, prompt-conditional, distributed across token positions, entangled with valence, implemented in attention or MLP subspaces not captured by final-token centroids, or simply too weakly expressed by these small instruction-tuned models.

The result also narrows the comparison with Wang et al. [9]. Their circuit claim concerns a 3B

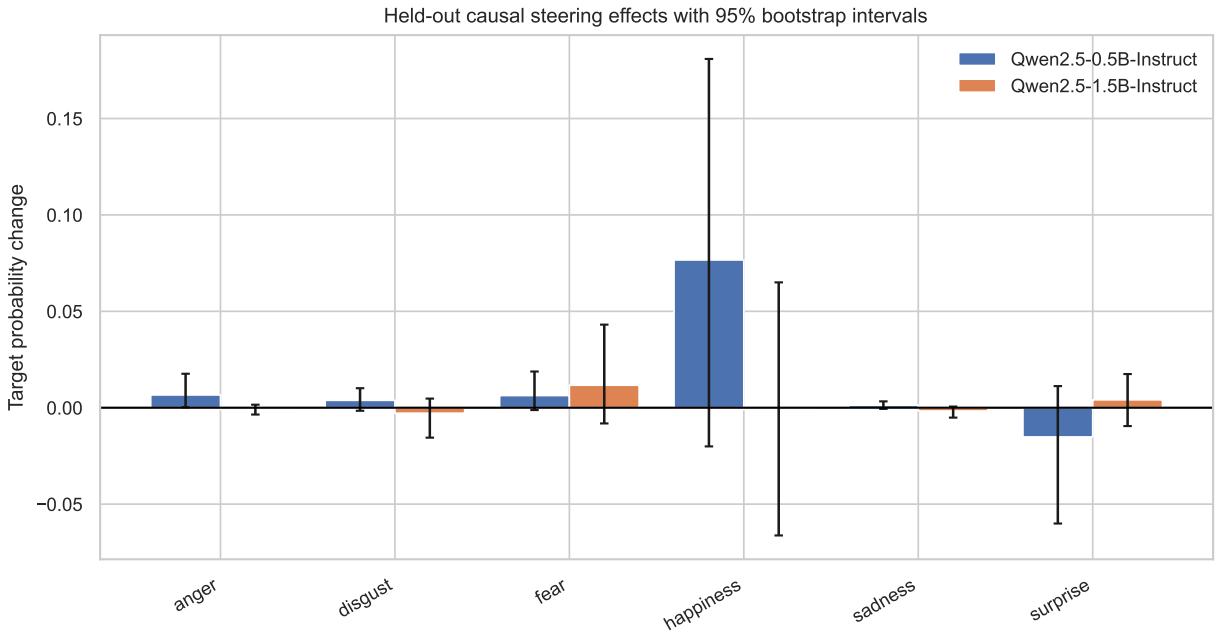

Figure 2: Target-direction effects relative to unsteered generation. Error bars are 95% scenario-bootstrap intervals.

Llama checkpoint, local neuron and head contributions, and integrated component interventions. Our data do not refute that result. Instead, they show that its existence cannot be inferred by extrapolating downward from high probe accuracy: no transition was detected at or below 1.5B under a reproducible version of the global-direction stage.

Why the 1.5B model may be harder to steer. All 1.5B development gains were negative despite superior representation accuracy. The selected late-layer window may encode emotion for classification while lying after the computation most useful for generation control. Alternatively, matched centroids may collapse distinct contextual modes more severely in the larger model. These are testable hypotheses, not post-hoc explanations: layer-by-layer causal maps, nonlinear subspaces, and component patching should be evaluated on new development data before another held-out test.

7 Limitations

First, the study contains only two checkpoints from one model family; it cannot locate a universal emergence threshold. The 3B extension was not run, and gated Llama weights prevented exact replication of the 1B target from Tak et al. [7]. Second, 36 held-out sufficiency skeletons and 18 necessity skeletons provide limited power for small, emotion-specific effects, although paired controls and 2,000 bootstrap resamples quantify this uncertainty.

Third, the output evaluator is an imperfect observational instrument. Its published model card reports 66% accuracy on a seven-class evaluation [3]; in our generated domain, balanced accuracy was 0.37–0.51 on explicit prompts, with a positive-class bias. Because every paired condition used the same evaluator, this does not create target effects by itself, but it can attenuate or distort them.

Human blinded annotation and multiple evaluators are important replication targets.

Fourth, direction estimation used explicitly instructed generated responses. Although within-skeleton centering removes scenario content, it cannot fully separate emotional style, instruction following, valence, and generic generation quality. Fifth, the neutral perplexity test intervened at all token positions and is deliberately stricter than the last-token decode-time intervention. It measures utility disruption but is not an exact simulation of generation-time exposure.

Finally, the criteria were intentionally stringent: H3 required every emotion to beat three comparisons after joint correction. This protects the term "emotion circuit" from isolated or selective effects but means that a partial mechanism would be reported as a failed family-level hypothesis. Per-emotion estimates are therefore provided rather than hidden behind the aggregate decision.

8 Reproducibility, Provenance, and Research Integrity

The complete local archive contains the prospectively frozen internal analysis plan, timestamped amendments, source commit identifiers, environment lockfile, model revision hashes, data and activation hashes, extraction code, intervention hooks, scored generation tables, tests, figures, and this manuscript. Prepared data exclude participant and demographic fields. All 2,412 sufficiency rows per model are retained, including control generations, and no empty output was discarded.

The public release contains the report, source code, tests, figures, aggregate statistics, and provenance record. Raw scored-generation tables are withheld from the public package because they reproduce text derived from upstream scenario datasets; they remain available in the local audit archive subject to upstream data and licensing constraints.

An initial float16 Metal run produced non-finite 1.5B activations. Those artifacts were invalidated before analysis, moved to a labeled `invalidated/float16` directory, and replaced by finite float32 extraction for both models. Probe artifacts record finite-coefficient and iteration diagnostics. The correction from the non-narrative validation CSV to a temporal round split occurred before model inference. The non-positive strength-selection amendment occurred after 1.5B development evaluation but before any 1.5B held-out intervention output.

AI-generation and human-role disclosure. OpenAI GPT-5.6 Sol, using a high-reasoning configuration through Codex, performed the literature retrieval, experimental design, software implementation, experiment execution, statistical analysis, figure production, validation workflow, and manuscript drafting. Josiah Wilson posed the initiating question, authorized use of local computing resources, observed the workflow, requested rigorous rather than simplified reporting, questioned the ethics of authorship, and authorized release under explicit AI attribution. He does not claim to understand or independently verify the technical mathematics and statistics. The work is therefore released as an unverified AI-generated research artifact, not as a conventional human-authored scholarly article. The repository retains sufficient code, outputs, amendments, and provenance for a qualified independent audit.

9 Conclusion

Across Qwen2.5 models at 0.5B and 1.5B parameters, emotion-related information was strongly and increasingly decodable, but matched linear directions were neither reliably sufficient nor necessary for emotional expression under controlled held-out interventions. The data therefore support the phrase *emotion-related representation*, not *emotion circuit*, for these models and methods. More broadly, the experiment demonstrates that scale-dependent improvements in interpretability probes

should not be read as evidence of a causal mechanism without direct, selective, and utility-preserving intervention tests.

References

[1] Marc E. Canby, Adam Davies, Chirag Rastogi, and Julia Hockenmaier. How reliable are causal probing interventions? *Proceedings of IJCNLP-AACL*, pages 857–878, 2025. doi: 10.18653/v1/2025.ijcnlp-long.47.

[2] Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021. doi: 10.1162/tacl_a_00359.

[3] Jochen Hartmann. Emotion english distilroberta-base. Hugging Face model card, 2022. URL `https://huggingface.co/j-hartmann/emotion-english-distilroberta-base`.

[4] John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of EMNLP-IJCNLP*, pages 2733–2743, 2019. doi: 10.18653/v1/D19-1275.

[5] Aishwarya Maheswaran and Maunendra Sankar Desarkar. A unified view on emotion representation in large language models. In *Proceedings of EACL*, pages 3589–3610, 2026. doi: 10.18653/v1/2026.eacl-long.165.

[6] Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2025. doi: 10.48550/arXiv.2412.15115.

[7] Ala N. Tak, Amin Banayeeanzade, Anahita Bolourani, Mina Kian, Robin Jia, and Jonathan Gratch. Mechanistic interpretability of emotion inference in large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 13090–13120, 2025. doi: 10.18653/v1/2025.findings-acl.679.

[8] Enrica Troiano, Laura Oberländer, and Roman Klinger. Dimensional modeling of emotions in text with appraisal theories: Corpus creation, annotation reliability, and prediction. *Computational Linguistics*, 49(1):1–72, 2023. doi: 10.1162/coli_a_00461.

[9] Chenxi Wang, Yixuan Zhang, Ruiji Yu, Yufei Zheng, Lang Gao, Zirui Song, Zixiang Xu, Gus Xia, Huishuai Zhang, Dongyan Zhao, and Xiuying Chen. Do llms "feel"? emotion circuits discovery and control. *arXiv preprint arXiv:2510.11328*, 2025. doi: 10.48550/arXiv.2510.11328.

[10] Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023. doi: 10.48550/arXiv.2310.01405.

A Development strength curves

B Direction geometry

Table 4: Mean development target-probability change by intervention strength.

Model	0.25	0.50	1.00	1.50	2.00
Qwen2.5-0.5B	0.0078	0.0114	0.0162	0.0344	0.0458
Qwen2.5-1.5B	-0.0099	-0.0110	-0.0256	-0.0097	-0.0201

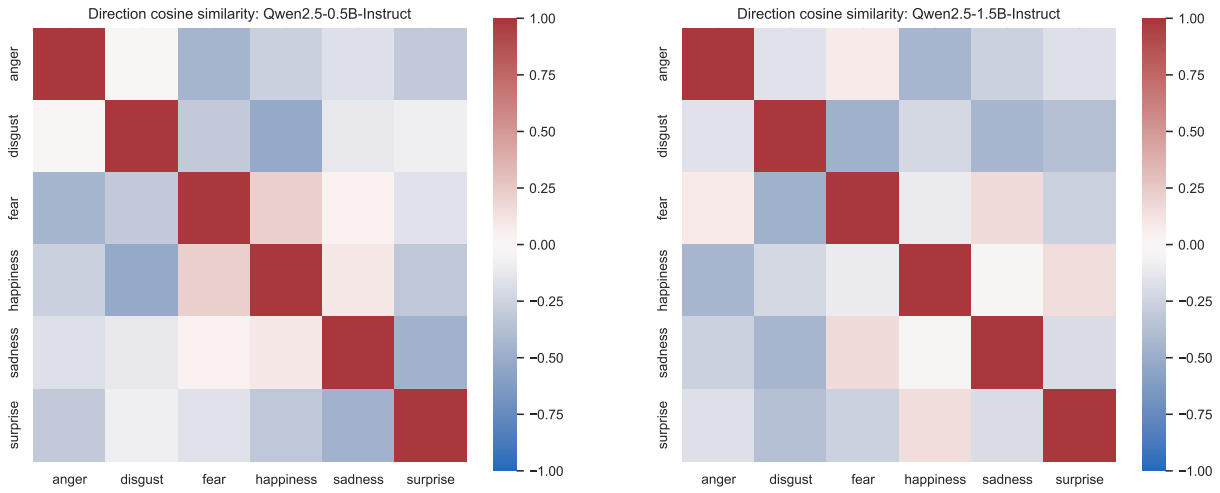

Figure 3: Cosine similarity among emotion directions at each model's selected probe layer. The directions are distinct but not mutually orthogonal.