Where Do Models Find Happiness? Emotion Vectors in Open-Source LLMs

Sinie van der Ben [1] **Raphaël Baur** [1] **Yannick Metz** [1] **Mennatallah El-Assady** [1]

Abstract

Recent work identified "emotion vectors" in Claude Sonnet 4.5, which are internal representations that encode emotion concepts, causally influence behavior, and exhibit geometry mirroring human psychological structure. We test the generality of these findings in two open-weight models, APERTUS-8B and GEMMA-4-E4B, extracting emotion contrast vectors across all layers, using two model-generated corpora. We recover valence geometry for both models, with peak PC1–valence correlations of $r = 0.76$ and $r = 0.83$, approaching the $r = 0.81$ reported for Claude. Beyond replication, we observe notable differences in how valence representations emerge across model depth. In GEMMA-4-E4B, valence is strongly encoded in early layers but collapses towards later layers, whereas APERTUS-8B exhibits the opposite pattern, with valence representations absent in early layers, but emerging at mid depths. Arousal encoding, in contrast, is sensitive to the extraction corpus: both models show stronger PC2–arousal alignment with Gemma-generated stories (r up to 0.45) than Apertus-generated ones ($r \leq 0.21$), suggesting arousal-relevant cues are unevenly distributed across generated corpora. We open-source our experiment code and dataset for reproducible investigation of emotion representations across language model architectures.

1. Introduction

As users interact with Large Language Models (LLM), they can encounter responses that appear emotionally reactive, such as expressing frustration when struggling with tasks or enthusiasm when helping users. Recent work by Sofroniew et al. (2026) moved beyond surface-level observations, identifying internal "emotion vectors" in Claude Sonnet 4.5. They identified 171 linear directions in activation space corresponding to emotion concepts, with correlational and potentially causal relations to model behaviour. Steering these vectors altered the model's preferences and increased rates of misaligned behaviors such as reward hacking and blackmail. The overall geometry of the emotion space mirrors human psychology, with principal components aligning to valence and arousal axes consistent with Russell's circumplex model (Russell, 1980). These findings raise key questions about generality: **(1)** Are emotion vectors specific to Claude's training, or a general property of language models' internal representations? **(2)** How does emotion geometry evolve across layers: Does it emerge suddenly or build up gradually? **(3)** How does the choice of story corpus affect extraction? These questions matter for interpretability and safety: If emotion representations are universal and robustly extractable, monitoring them could provide early warnings of misaligned internal states across different models. We address these questions by replicating and extending emotion vector analysis in two open-weight models: APERTUS-8B (Hernández-Cano et al., 2025), with fully open weights, training data, and code, and GEMMA-4-E4B (DeepMind, 2026), a recently released open-source model, both chosen for their relatively small size. For each model, we extract emotion contrast vectors across multiple layers using two story corpora—one generated by APERTUS-8B and one by GEMMA-4-E4B —to separate model-intrinsic geometry from corpus-dependent extraction artifacts. Additional related work is provided in Appendix A. We release our code publicly[1].

- **Replication of key findings.** We recover valence geometry in both APERTUS-8B and GEMMA-4-E4B, with the highest PC1–valence correlations of r=0.76 and r=0.83 respectively, demonstrating that emotion vectors generalize beyond Claude to open-weight models across different architectures.
- **Divergent Emergence.** Models differ substantially in when valence structure emerges: GEMMA-4-E4B peaks early (layer 16) then fades, while APERTUS-8B builds progressively across depth, stabilizing around layer 20. Cross-layer CKA analysis shows a phase transition in APERTUS-8B that is absent in Gemma.

[1]Department of Computer Science, ETH, Zurich, Switzerland. Correspondence to: Sinie van der Ben <vandersi@ethz.ch>.

Published at the Mechanistic Interpretability Workshop at ICML 2026. Copyright 2026 by the author(s).

[1]https://github.com/sinievanderben/emotion_experiment

- **Corpus-dependent arousal.** Arousal encoding is sensitive to story corpus: both models show substantially stronger PC2–arousal alignment when using Gemma-generated stories (r up to 0.45) than Apertus-generated stories (r $\leq$ 0.17)

2. Methods

2.1. Dataset

We generated **two** synthetic emotion-story datasets, following Sofroniew et al. (2026), with 9 stories for each of 171 emotions. For each emotion, we prompted APERTUS-8B and GEMMA-4-E4B to write short stories in which characters experience the target emotion without naming it, using a similar prompt to Sofroniew et al. (2026). This produced 1,539 stories across emotions (Table 1), plus 40 neutral stories from the same model. The 40 neutral texts form a single fixed set shared by all 171 emotions, since we compute the confound subspace once per layer and project every emotion vector through the same operation. The emotion concepts span the valence-arousal space.

We treat the story corpora as independent variable. By running each model on both Apertus-generated and Gemma-generated stories, we intend to disentangle the emotion findings from corpus-dependent extraction artifacts. No previous work has tested story influence before.

2.2. Model

We analyzed two open-weight language models: APERTUS-8B Instruct, a 32-layer transformer, and GEMMA-4-E4B, a 42-layer transformer. Both models are instruction-tuned and comparable in scale to enable cross-model comparison of emotion representations. More details on both models can be found in Appendix B.2.

2.3. Contrast Vector Extraction

Following Sofroniew et al. (2026), we construct one activation vector $\mathbf{v}_e^{(l)}$ per emotion e and layer l. Since these vectors capture general linguistic structure, we apply a two-step procedure to isolate the emotion-specific component.

First, for each emotion, we perform a forward pass on the corresponding nine stories and cache the residual stream activations at each layer, giving a tensor of shape $(\#\text{tokens}, d_{\text{model}})$ per layer. Averaging these activations across tokens and stories yields one raw vector $\mathbf{u}_e \in \mathbb{R}^{d_{\text{model}}}$ per emotion and layer, which still mixes emotion-specific and general linguistic features.

Second, we project out non-emotion-specific components. To characterize the emotion-agnostic subspace, we collect mean residual activations from the 40 neutral stories, producing a $(40, d_{\text{model}})$ matrix per layer. PCA on this matrix yields a basis for the subspace; we retain the top K components that together explain 50% of the variance. To isolate the emotion-specific component, we subtract from each emotion vector its projection onto the neutral subspace to get the contrast vector $\mathbf{v}_e$:

$$\mathbf{v}_e = \mathbf{u}_e - \sum_{k=1}^{K} (\mathbf{u}_e \cdot \mathbf{p}_k)\mathbf{p}_k$$

For APERTUS-8B, we extracted vectors from layers *1–31*, and for GEMMA-4-E4B, from layers *1–40*. Stacking these vectors across all $|E|$ emotions yields the matrix $V^{(l)} \in \mathbb{R}^{|E| \times d_{\text{model}}}$ at layer l, on which we perform the analyses.

2.4. Analysis

PCA and Valence-Arousal Correlation We applied PCA to the emotion contrast matrix $V^{(l)}$ at each layer and correlated the first two principal components (PC1, PC2) with human valence and arousal ratings from the NRC Valence–Arousal–Dominance Lexicon (Mohammad, 2018), following (Sofroniew et al., 2026). We report Pearson r and corresponding p-values.

Cross-layer Representational Similarity with CKA We computed linear Centered Kernel Alignment (Kornblith et al., 2019) between $V^{(l)}$ for all layer pairs within each model and story condition. CKA values near 1 indicate similar representational geometry, while values near 0 indicate orthogonal structure. Because CKA is invariant to orientation in latent space, it is well suited for this comparison and allowed us to quantify how emotion geometry evolves through the network.

Valence direction stability Lastly, we identified the valence direction at each layer as the vector most correlated with human valence ratings (using PC1 when this correlation is significant), then computed cosine similarity between these directions across layers to test whether the same subspace encodes valence at different depths.

3. Results

3.1. Valence Replicates Across Models and Corpora

The first principal component of the emotion contrast matrix aligns with human valence ratings in both models, replicating the main result of (Sofroniew et al., 2026). Figure 1 shows PC1–valence correlations across fractional layer depth for all four model$\times$corpus conditions; per-layer values are reported in Table 3.

Peak correlations. All model$\times$corpus combinations reach a peak between $r = 0.72$ and 0.83. APERTUS-8B peaks at $r = 0.72$ (layer 23, Apertus stories) and $r = 0.76$ (layer 31, Gemma stories); GEMMA-4-E4B peaks at $r =$

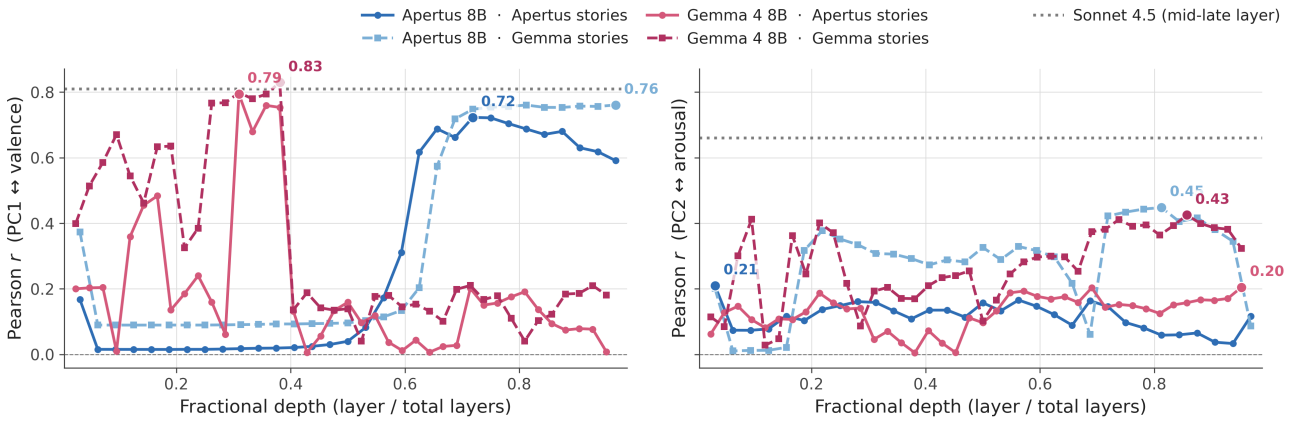


Figure 1. Pearson correlation between the top two PCs of the emotion-vector space and human valence (left, PC1) and arousal (right, PC2) across fractional layer depth, for Apertus 8B and GEMMA-4-E4B probed on Apertus- and Gemma-generated stories (four conditions). Hue = model (blue = Apertus, red = Gemma); line style = story source (solid = Apertus, dashed = Gemma). Dotted gray lines mark the Sonnet 4.5 reference at a mid-late layer ($r = 0.81$ valence, $r = 0.66$ arousal; (Sofroniew et al., 2026)).

0.79 (layer 13, Apertus stories) and $r = 0.83$ (layer 16, Gemma stories). All peaks are significant ($p < 10^{-3}$) and approach or exceed the Sonnet 4.5 reference of $r = 0.81$.

Valence Across Network Depth Both models reach similar r-value peaks with opposite depth profiles. APERTUS-8B shows *abrupt late emergence*: PC1–valence correlation is near zero through fractional depth ≈ 0.5 (layer 17/18), then rises sharply, becoming significant at layer 18 ($r = 0.17$, $p < 0.05$) and exceeding $r = 0.60$ at layer 21 ($\approx$63% depth) under both story conditions.

GEMMA-4-E4B instead shows *early encoding followed by collapse*: for Apertus stories, valence peaks at layer 16 ($\approx$38% depth), then falls near zero by layer 18, with only partial recovery ($r \approx 0.18$–0.20) in the final layers. For Gemma stories, the peak comes later and both pre- and post-peak values are higher. The Sonnet 4.5 reference peaks in the mid-late range, indicating that APERTUS-8B follows a similar pattern.

Representation space vs. valence-axis stability To interpret valence trajectories, we examine **(i)** whole-space representational similarity via linear CKA and **(ii)** cosine alignment of the layer-wise valence direction for each model–corpus combination.

APERTUS-8B (Figs. 2, 3) shows three CKA phases: layers 2–11 form a flat plateau (CKA ≈ 1); layers 12–21 form a transition band with off-diagonal decay (minimum 0.33 on Apertus stories, 0.58 on Gemma stories); and layers 22–31 form a second plateau. This transition aligns with the rise of PC1–valence correlation, suggesting a representational reorganization. GEMMA-4-E4B (Figs. 4, 5) instead shows a smooth gradient across all 40 layers with no sharp transition and CKA ≥ 0.73 between any pair. The collapse of Gemma's valence correlation around layer 18 therefore cannot stem from a global reorganization, as the geometry remains approximately stable through the collapse.

The valence-direction cosine matrices show what changes. For APERTUS-8B on its own stories (Fig. 6), no off-diagonal cell exceeds $|0.49|$, which can indicate that the recovered direction is noise across layers. On Gemma stories (Fig. 7), early layers (2–11) form a coherent block with cosines 0.35–0.57 before becoming noisy in later layers. For GEMMA-4-E4B on Gemma stories (Fig. 10), there are two positive blocks (layers 2–8 and 9–14) and a late block (28–40), with adjacent-layer cosines up to ± 0.55. On Apertus stories (Fig. 11) this structure is less pronounced. Because CKA matrices are similar across corpora, the emotion representational space is corpus-invariant. However, the recovered valence axis depends on the input corpus, with Gemma stories yielding cleaner valence directions in both models. Thus, valence is recoverable in both, but not encoded along a consistent axis across depth.

PCA cluster separation at peak layers PCA projections at each model's peak layer (Figs. 14, 15) show emotion clustering and a clear corpus effect. PC1–valence correlations are similar across story conditions (APERTUS-8B L23: 0.72 vs. 0.75; GEMMA-4-E4B L13: 0.79 vs. 0.80), but clusters are more clearly separated for Gemma stories, with positive and negative emotions forming denser groups.

3.2. Arousal Encoding

PC2–arousal correlations are generally weaker than PC1–valence and depend strongly on the story corpus (Fig. 1, right; Table 4). On Apertus stories, both models peak below $r = 0.21$ (APERTUS-8B: $r = 0.17$ at layer 18; GEMMA-4-E4B: $r = 0.21$ at layer 40). On Gemma sto-

ries, both models reach $r > 0.40$ (APERTUS-8B: $r = 0.45$ at layer 26; GEMMA-4-E4B: $r = 0.41$ at layer 31, both $p < 10^{-8}$). Possibly, Gemma-generated stories contain more arousal-discriminative linguistic content. We leave a corpus-content analysis to future work.

4. Discussion

Main Research Questions Our results address the three questions raised in the introduction. **(1)** *Emotion vectors are not specific to Claude's training*. We recover a valence axis of similar strength in two architecturally distinct open-weight models, with peak correlations matching ($r = 0.83$ for GEMMA-4-E4B) or approaching ($r = 0.76$ for APERTUS-8B) the $r = 0.81$ reported for Claude Sonnet 4.5. **(2)** *Emergence is not uniform across models*. APERTUS-8B builds valence alignment abruptly in the second half of the network, while GEMMA-4-E4B encodes it early and then loses it mid-network. **(3)** *The story corpus affects extraction*. This is especially clear for arousal: Gemma-generated stories yield correlations more than twice as large as Apertus-generated stories in both probed models.

Different paths to the same geometry. GEMMA-4-E4B and APERTUS-8B reach similar peak valence correlations ($r \approx 0.76$–0.83) via different layer-wise trajectories. GEMMA-4-E4B encodes valence in earlier layers before it degrades in later layers, while APERTUS-8B develops it sharply across mid-to-upper layers. We have not yet explored the possible attribution of this to architecture, training data, or post-training, since the models differ in all three. Our results show that similar peak valence correlations can hide substantial differences in *where* and *how* valence is computed.

Stable representation space, unstable axis The representational space (CKA) and valence-axis stability disconnect. In GEMMA-4-E4B, the space remains similar across layers even where the PC1–valence correlation collapses, so valence information is preserved. In APERTUS-8B, the valence axis is relatively unstable across layers despite a late high plateau of valence–PC1 correlation. Thus, representational similarity between layers does not guarantee a shared valence direction.

The arousal gap and corpus dependence Arousal shows the weakest replication, but the story-condition analysis suggests that this may be attributed to our methodological choices. With Gemma-generated stories, arousal correlations in both models rise (from $r \leq 0.21$ to $r \geq 0.43$), partially closing the gap with the original result ($r = 0.66$). Because Gemma stories improve arousal extraction in *both* models, the effect likely reflects corpus properties rather than model–story matching. Since it appears in both models, this rules out the simple confound that each model encodes only its own corpus well. We hypothesize that Gemma has the ability to generate stories with greater variation in narrative intensity and physiological arousal cues, so corpus choice for eliciting emotion contrasts is a substantive methodological factor, not an implementation detail. We leave verification to future work.

4.1. Limitations

Several limitations warrant mention. The first, the original study (Sofroniew et al., 2026) did not release code, so our implementation is reconstructed from the methods they described. Subtle methodological differences may therefore contribute to numerical differences. Second, our analysis covers two open-weight models from two families. Broader cross-architecture comparisons would strengthen claims about how general the valence-pattern is, and whether the trajectory differences generalize to other model families. Third, the corpora we probe are themselves model-generated, which means we cannot fully separate properties of the distributions it produces. A fully model-independent stimulus set would be a stronger control.

4.2. Future Work

Several directions follow from our findings and limitations. The most direct is causal validation: steering model outputs at peak-correlation layers along the recovered valence direction would test whether the representational structure we identify is actually used by the model. Related, the cross-layer rotations of the valence axis raises the question whether steering vectors derived at one layer remain effective when applied to another, even within regions of overall stable space. Cross-layer feature tracking using sparse autoencoders could further reveal whether the same interpretable features carry emotion information across the depth ranges we identify, or whether different layers encode emotions through different feature combinations. Finally, extending this analysis to multi-modal models could test whether the valence axis is preserved across modalities.

5. Conclusion

We replicate Anthropic's emotion findings in two open-weight models, achieving valence correlations of $r = 0.83$ (GEMMA-4-E4B) and $r = 0.76$ (APERTUS-8B). Cross-layer analysis reveals divergent developmental trajectories: GEMMA-4-E4B encodes valence in early layers while APERTUS-8B builds it progressively through late layers. These results suggest that similar representations can arise from different computational paths, with implications for layer selection in interpretability work and targeted steering interventions.

6. Acknowledgement

We would like to thank Steffen Holter for his valuable feedback on the final stages of the paper. Raphaël Baur is supported by the ETH AI Center Fellowship. This work was supported under project ID a127 as part of the Swiss AI Initiative, through a grant from the ETH Domain and computational resources provided by the Swiss National Supercomputing Centre (CSCS) under the Alps infrastructure.

References

Arditi, A., Obeso, O. B., Syed, A., Paleka, D., Rimsky, N., Gurnee, W., and Nanda, N. Refusal in language models is mediated by a single direction. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=pH3XAQME6c.

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., Askell, A., et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023.

Cheng, E., Doimo, D., Kervadec, C., Macocco, I., Yu, L., Laio, A., and Baroni, M. Emergence of a high-dimensional abstraction phase in language transformers. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=0fD3iIBhlV.

Choi, B. J. and Weber, M. Latent structure of affective representations in large language models, 2026. URL https://arxiv.org/abs/2604.07382.

Cunningham, H., Ewart, A., Riggs, L., Huben, R., and Sharkey, L. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023.

DeepMind, G. Gemma 4: Expanding the gemmaverse with apache 2.0, 2026. URL https://opensource.googleblog.com/2026/03/gemma-4-expanding-the-gemmaverse-with-apache-20.html. Accessed: 2026-04-28.

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

Hernández-Cano, A., Hägele, A., Huang, A. H., Romanou, A., Solergibert, A.-J., Pasztor, B., Messmer, B., Garbaya, D., Ďurech, E. F., Hakimi, I., Giraldo, J. G., Ismayilzada, M., Foroutan, N., Moalla, S., Chen, T., Sabolčec, V., Xu, Y., Aerni, M., AlKhamissi, B., Marinas, I. A., Amani, M. H., Ansaripour, M., Badanin, I., Benoit, H., Boros, E., Browning, N., Bösch, F., Böther, M., Canova, N., Challier, C., Charmillot, C., Coles, J., Deriu, J., Devos, A., Drescher, L., Dzenhaliou, D., Ehrmann, M., Fan, D., Fan, S., Gao, S., Gila, M., Grandury, M., Hashemi, D., Hoyle, A., Jiang, J., Klein, M., Kucharavy, A., Kucherenko, A., Lübeck, F., Machacek, R., Manitaras, T., Marfurt, A., Matoba, K., Matrenok, S., Mendoncça, H., Mohamed, F. R., Montariol, S., Mouchel, L., Najem-Meyer, S., Ni, J., Oliva, G., Pagliardini, M., Palme, E., Panferov, A., Paoletti, L., Passerini, M., Pavlov, I., Poiroux, A., Ponkshe, K., Ranchin, N., Rando, J., Sauser, M., Saydaliev, J., Sayfiddinov, M. A., Schneider, M., Schuppli, S., Scialanga, M., Semenov, A., Shridhar, K., Singhal, R., Sotnikova, A., Sternfeld, A., Tarun, A. K., Teiletche, P., Vamvas, J., Yao, X., Ilic, H. Z. A., Klimovic, A., Krause, A., Gulcehre, C., Rosenthal, D., Ash, E., Tramèr, F., VandeVondele, J., Veraldi, L., Rajman, M., Schulthess, T., Hoefler, T., Bosselut, A., Jaggi, M., and Schlag, I. Apertus: Democratizing Open and Compliant LLMs for Global Language Environments. https://arxiv.org/abs/2509.14233, 2025.

Kornblith, S., Norouzi, M., Lee, H., and Hinton, G. Similarity of neural network representations revisited. In *International conference on machine learning*, pp. 3519–3529. PMlR, 2019.

Marks, S. and Tegmark, M. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets, 2024. URL https://arxiv.org/abs/2310.06824.

Mikolov, T., Yih, W.-t., and Zweig, G. Linguistic regularities in continuous space word representations. In Vanderwende, L., Daumé III, H., and Kirchhoff, K. (eds.), *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 746–751, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL https://aclanthology.org/N13-1090/.

Mohammad, S. M. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In *Proceedings of ACL*, 2018.

Park, K., Choe, Y. J., and Veitch, V. The linear representation hypothesis and the geometry of large language models. In *Causal Representation Learning Workshop at NeurIPS 2023*, 2023. URL https://openreview.net/forum?id=T0PoOJg8cK.

Radford, A., Jozefowicz, R., and Sutskever, I. Learning to generate reviews and discovering sentiment, 2017. URL https://arxiv.org/abs/1704.01444.

Rimsky, N., Gabrieli, N., Schulz, J., Tong, M., Hubinger, E., and Turner, A. Steering llama 2 via contrastive activation addition. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.828. URL https://aclanthology.org/2024.acl-long.828/.

Russell, J. A. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.

Sofroniew, N., Kauvar, I., Saunders, W., Chen, R., Henighan, T., Hydrie, S., Citro, C., Pearce, A., Tarng, J., Gurnee, W., Batson, J., Zimmerman, S., Rivoire, K., Fish, K., Olah, C., and Lindsey, J. Emotion concepts and their function in a large language model. *Transformer Circuits Thread*, 2026. URL https://transformer-circuits.pub/2026/emotions/index.html.

Sun, L., Yan, L., Lu, X., Lee, A., Zhang, J., and Shao, J. Valence-arousal subspace in llms: Circular emotion geometry and multi-behavioral control, 2026. URL https://arxiv.org/abs/2604.03147.

Tigges, C., Hollinsworth, O. J., Geiger, A., and Nanda, N. Language models linearly represent sentiment. In *ICML 2024 Workshop on Mechanistic Interpretability*, 2024. URL https://openreview.net/forum?id=Xsf6dOOMMc.

Valeriani, L., Doimo, D., Cuturello, F., Laio, A., ansuini, A., and Cazzaniga, A. The geometry of hidden representations of large transformer models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=cCYvakU5Ek.

A. Related Work

Linear representations in LLMs. The linear representation hypothesis holds that high-level concepts are encoded as directions in activation space (Mikolov et al., 2013; Elhage et al., 2022; Park et al., 2023). Tigges et al. (2024) demonstrated this for sentiment, finding a single direction captures positive-negative valence across tasks. Subsequent work extended linear representations to truth (Marks & Tegmark, 2024), refusal (Arditi et al., 2024), and behavioral tendencies (Rimsky et al., 2024). Sparse autoencoders can extract directions at scale, decomposing polysemantic activations into interpretable features (Bricken et al., 2023; Cunningham et al., 2023).

Emotion in language models. Early work identified a "sentiment neuron" in LSTMs (Radford et al., 2017), though later analysis suggested emotional content is distributed across many neurons. (Sofroniew et al., 2026) provide a comprehensive analysis, extracting 171 emotion vectors from Claude Sonnet 4.5 and demonstrating causal influence on behavior. They found emotion geometry mirrors human psychological structure, with valence and arousal as principal axes. Concurrent work extends this to other models: (Sun et al., 2026) identify a valence-arousal subspace in Llama and Qwen with circumplex-consistent circular geometry, where steering along VA axes controls refusal and sycophancy. (Choi & Weber, 2026) find coherent affective representations in Gemma-2, Mistral, and LLaMA with modest nonlinear global structure. We build on Sofroniew et al. (2026), testing generalization across architectures and the role of extraction methodology.

Cross-layer geometry. Transformer representations evolve across layers in characteristic ways. Valeriani et al. (2023) found intrinsic dimension expands then contracts, with semantics concentrated at intermediate depths. (Cheng et al., 2025) identified a "high-dimensional abstraction phase" where representations peak in complexity before simplifying toward outputs.

B. Story Dataset

The emotion story datasets was generated using APERTUS-8B and GEMMA-4-E4B, following a methodology similar to Anthropic's emotion vectors work (Sofroniew et al., 2026). The 171 emotions were copied from their work. Stories were designed to convey emotions implicitly, such as never naming the target emotion directly, but instead relying instead on character actions, physical sensations, dialogue, and situational context. The prompts used were also similar, to introduce as little methodological confound as possible.

B.1. Dataset statistics

Table 1. Emotion story dataset statistics by corpus. Apertus stories were deduplicated to match the uniform 9-stories-per-topic structure of the Gemma corpus.

Statistic	**Apertus stories**	**Gemma stories**
Generator model	APERTUS-8B-Instruct-2509	GEMMA-4-E4B-it
Total emotions	171	171
Unique topics	100	100
Stories per emotion	9	9
Total stories	1,539	1,539
Mean story length	$\sim$215 words	$\sim$144 words
Story length range	65–665 words	81–298 words

B.2. Activation collection

Residual stream activations were collected from APERTUS-8B-Instruct at multiple transformer layers (Table 2). The model can be found through HuggingFace: `swiss-ai/APERTUS-8B-Instruct-2509`.

For Gemma, different layers were picked to collect activations from (Table 2). The model was also accessed through HuggingFace: `google/GEMMA-4-E4B-it`.

Table 2. Activation extraction configuration for both models.

Parameter	**APERTUS-8B**	**GEMMA-4-E4B 8B**
Model	APERTUS-8B-Instruct-2509	GEMMA-4-E4B
Layers collected	1-31	1-40
Hook location	`resid_post`	`resid_post`
Batch size	8 sequences	8 sequences
Max sequence length	1024 tokens	1024 tokens
Dataset	Pile (uncopyrighted), train split	Pile (uncopyrighted), train split

C. Additional Results

C.1. Principal Component Valence

Table 3. PC1–valence (Pearson r) across layers and story conditions. Bold indicates the peak layer per model–condition pair. $^{\dagger}p < 0.05$; $^{\ddagger}p < 0.01$; $^{*}p < 0.001$.

	APERTUS-8B			GEMMA-4-E4B 8B	
	APERTUS-8B *stories*	GEMMA-4-E4B *stories*		APERTUS-8B *stories*	GEMMA-4-E4B *stories*
Layer	r	r	**Layer**	r	r
1	$0.1675^{\dagger}$	0.3737^{*}	1	$0.2000^{\ddagger}$	0.3986^{*}
2	0.0154	0.0895	2	$0.2030^{\ddagger}$	0.5138^{*}
3	0.0152	0.0894	3	$0.2040^{\ddagger}$	0.5857^{*}
4	0.0152	0.0894	4	0.0098	0.6708^{*}
5	0.0151	0.0894	5	0.3592^{*}	0.5443^{*}
6	0.0151	0.0895	6	0.4550^{*}	0.4619^{*}
7	0.0154	0.0895	7	0.4835^{*}	0.6333^{*}
8	0.0153	0.0894	8	0.1359	0.6352^{*}
9	0.0163	0.0900	9	$0.1851^{\dagger}$	0.3248^{*}
10	0.0176	0.0908	10	$0.2401^{\ddagger}$	0.3851^{*}
11	0.0188	0.0916	11	$0.1598^{\dagger}$	0.7664^{*}
12	0.0193	0.0928	12	0.0614	0.7675^{*}
13	0.0209	0.0928	13	$\mathbf{0.7940}^{*}$	0.7984^{*}
14	0.0246	0.0938	14	0.6797^{*}	0.7795^{*}
15	0.0302	0.0946	15	0.7595^{*}	0.7938^{*}
16	0.0400	0.0959	16	0.7533^{*}	$\mathbf{0.8296}^{*}$
17	0.0821	0.1034	17	0.1318	0.1360
18	$0.1732^{\dagger}$	0.1141	18	0.0060	$0.1880^{\dagger}$
19	0.3112^{*}	0.1343	19	0.0564	0.1418
20	0.6169^{*}	$0.2034^{\ddagger}$	20	0.1337	0.1350
21	0.6882^{*}	0.5739^{*}	21	$0.1591^{\dagger}$	0.1393
22	0.6616^{*}	0.7181^{*}	22	0.0968	0.0403
23	$\mathbf{0.7230}^{*}$	0.7478^{*}	23	0.1187	$0.1769^{\dagger}$
24	0.7216^{*}	0.7539^{*}	24	0.0370	$0.1796^{\dagger}$
25	0.7040^{*}	0.7561^{*}	25	0.0117	0.1446
26	0.6875^{*}	0.7606^{*}	26	0.0436	$0.1536^{\dagger}$
27	0.6710^{*}	0.7535^{*}	27	0.0067	0.1319
28	0.6803^{*}	0.7533^{*}	28	0.0242	0.1015
29	0.6302^{*}	0.7573^{*}	29	0.0277	$0.1986^{\ddagger}$
30	0.6183^{*}	0.7564^{*}	30	$0.2014^{\ddagger}$	$0.2101^{\ddagger}$
31	0.5910^{*}	$\mathbf{0.7608}^{*}$	31	$0.1505^{\dagger}$	$0.1672^{\dagger}$
32	—	—	32	$0.1567^{\dagger}$	$0.1793^{\dagger}$
33	—	—	33	$0.1754^{\dagger}$	0.1105
34	—	—	34	$0.1910^{\dagger}$	0.0408
35	—	—	35	0.1366	0.1034
36	—	—	36	0.0939	0.1222
37	—	—	37	0.0743	$0.1841^{\dagger}$
38	—	—	38	0.0786	$0.1861^{\dagger}$
39	—	—	39	0.0769	$0.2099^{\ddagger}$
40	—	—	40	0.0081	$0.1810^{\dagger}$

C.2. Principal Component Arousal

Table 4. PC2–arousal (Pearson r) across layers and story conditions. Bold indicates the peak layer per model–condition pair. $^{\dagger}p < 0.05$; $^{\ddagger}p < 0.01$; $^{*}p < 0.001$.

	APERTUS-8B			GEMMA-4-E4B	
Layer	APERTUS-8B *stories* r	GEMMA-4-E4B *stories* r	**Layer**	APERTUS-8B *stories* r	GEMMA-4-E4B *stories* r
1	**0.2093**‡	0.2012‡	1	0.0624	0.1141
2	0.0737	0.0105	2	0.1273	0.0845
3	0.0735	0.0122	3	0.1472†	0.3005*
4	0.0778	0.0124	4	0.1060	0.4120*
5	0.1168	0.0214	5	0.0815	0.0282
6	0.1032	0.3171*	6	0.1086	0.0476
7	0.1370	0.3772*	7	0.1064	0.3625*
8	0.1488†	0.3520*	8	0.1298	0.2450‡
9	0.1610†	0.3345*	9	0.1867†	0.4008*
10	0.1585†	0.3076*	10	0.1574†	0.3710*
11	0.1328	0.3060*	11	0.1384	0.2168‡
12	0.1085	0.2923*	12	0.1414	0.0872
13	0.1350	0.2731*	13	0.0459	0.1928†
14	0.1351	0.2883*	14	0.0707	0.2045‡
15	0.1137	0.2818*	15	0.0360	0.1709†
16	0.1573†	0.3262*	16	0.0048	0.1700†
17	0.1329	0.2890*	17	0.0726	0.2094‡
18	0.1657†	0.3290*	18	0.0345	0.2341‡
19	0.1469†	0.3166*	19	0.0054	0.2407‡
20	0.1211	0.2974*	20	0.1114	0.2547*
21	0.0891	0.2167‡	21	0.0965	0.1018
22	0.1637†	0.0607	22	0.1325	0.1660†
23	0.1459†	0.4222*	23	0.1879†	0.2452‡
24	0.0972	0.4341*	24	0.1931†	0.2819*
25	0.0810	0.4432*	25	0.1769†	0.2970*
26	0.0586	**0.4480***	26	0.1690†	0.2997*
27	0.0597	0.4049*	27	0.1755†	0.2979*
28	0.0654	0.4165*	28	0.1585†	0.2538*
29	0.0371	0.3804*	29	0.2030‡	0.3741*
30	0.0331	0.3449*	30	0.1437	0.3816*
31	0.1165	0.0866	31	0.1524†	0.4115*
32	—	—	32	0.1487	0.3904*
33	—	—	33	0.1393	0.3950*
34	—	—	34	0.1248	0.3638*
35	—	—	35	0.1509†	0.3926*
36	—	—	36	0.1577†	**0.4251***
37	—	—	37	0.1665†	0.3989*
38	—	—	38	0.1650†	0.3863*
39	—	—	39	0.1706†	0.3818*
40	—	—	40	**0.2047**‡	0.3234*

C.3. CKA Figures

CKA is a measure of how similar the emotion space is between 2 layers. The diagonal is always 1, which is a layer compared to itself.

- CKA close to 1: spatial arrangement of emotion vectors between 2 layers is nearly identical.
- CKA close to 0: spatial arrangement has changed substantially between 2 layers

So each cell answers: does the model organize emotions in the same way at layer A as in layer B? The higher the value, the more similar.

C.3.1. APERTUS-8B CKA RESULTS

Figure 2. **APERTUS-8B** CKA results on Apertus stories

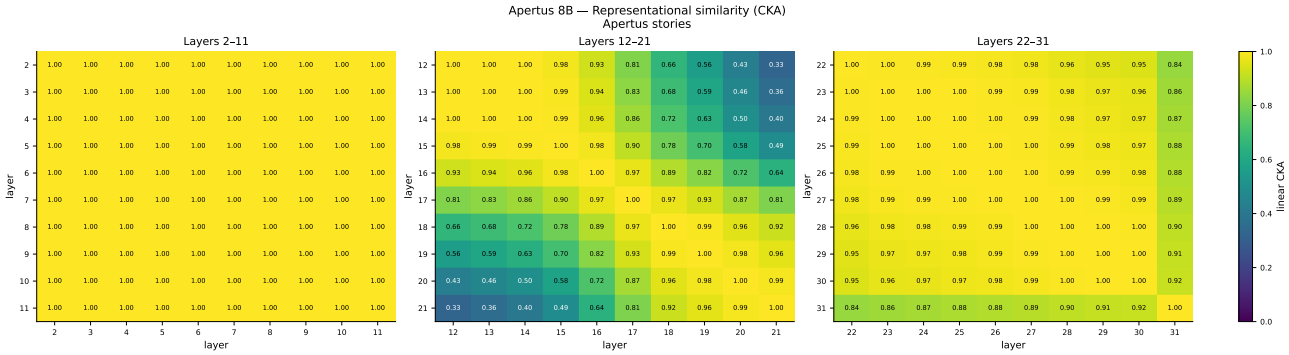
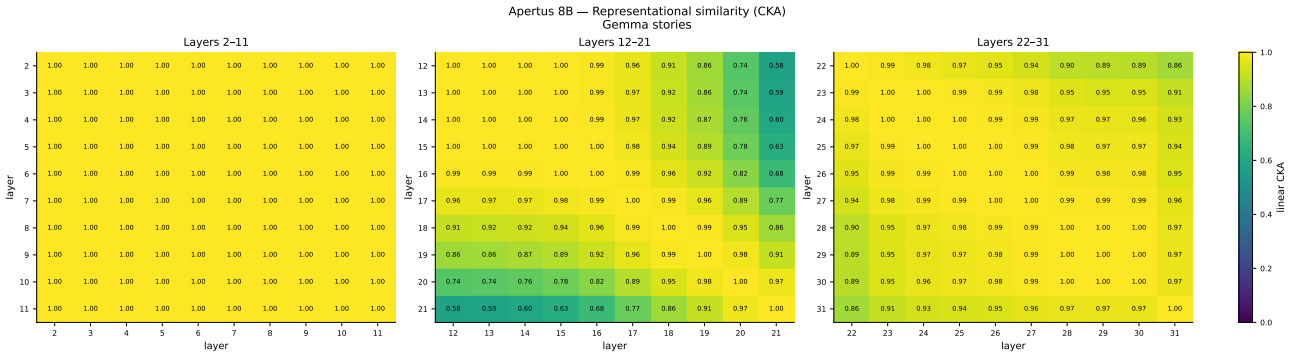


Figure 3. **APERTUS-8B** CKA values on Gemma stories



C.3.2. GEMMA-4-E4B CKA RESULTS

Figure 4. **GEMMA-4-E4B** CKA values on Gemma stories

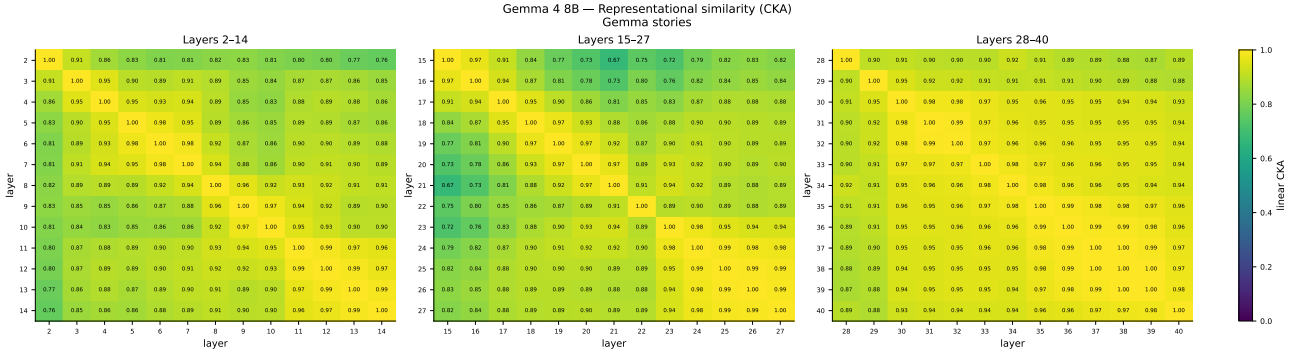
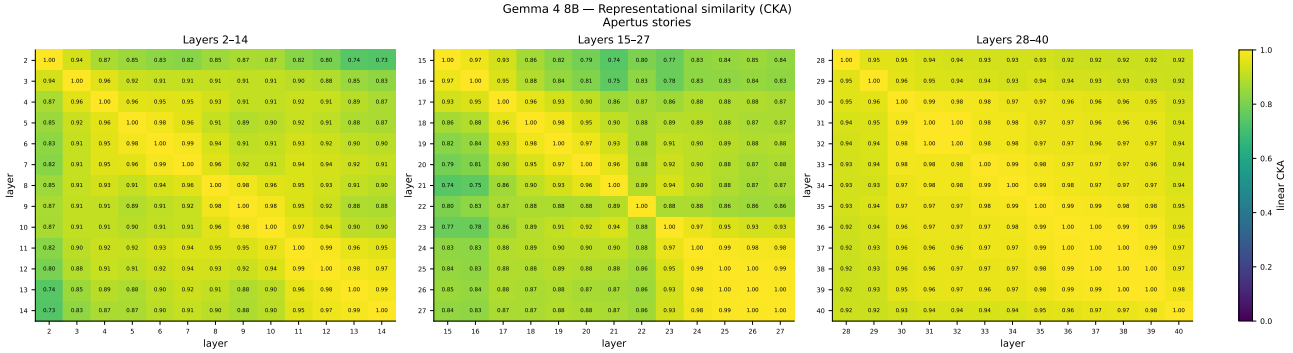


Figure 5. **GEMMA-4-E4B** CKA values on Apertus stories



C.4. Valence Direction Alignment

Each cell shows the cosine similarity between the valence direction vectors at 2 layers. The valence direction is the axis in activation space that best predicts the emotion valence.

- **Cosine similarity close to 1**. Valence axis points in the same direction in both layers, consistent positive axis.
- **Cosine similarity close to 0**. Valence axes are orthogonal, they've rotated completely.
- **Cosine similarity close to -1**. The axis has flipped direction.

CKA provides information about the whole space, while the cosine similarity specifically shows whether the valence axis is stable. A predominantly blue matrix would indicate that the model has a persistent stable direction to represent positive vs. negative emotions across many layers.

The valence direction stability line plot shows the cosine similarity between 2 adjacent layers. It has a similar interpretation as the values in the panel, but only for adjacent layers. The interpretation can be slightly different, because it shows if the valence axis points in the same direction from layer to layer. A dip reveals a specific transition, where the model changes how it encodes valence.

C.4.1. APERTUS-8B VALENCE ALIGNMENT

Figure 6. **APERTUS-8B** validation on Apertus stories

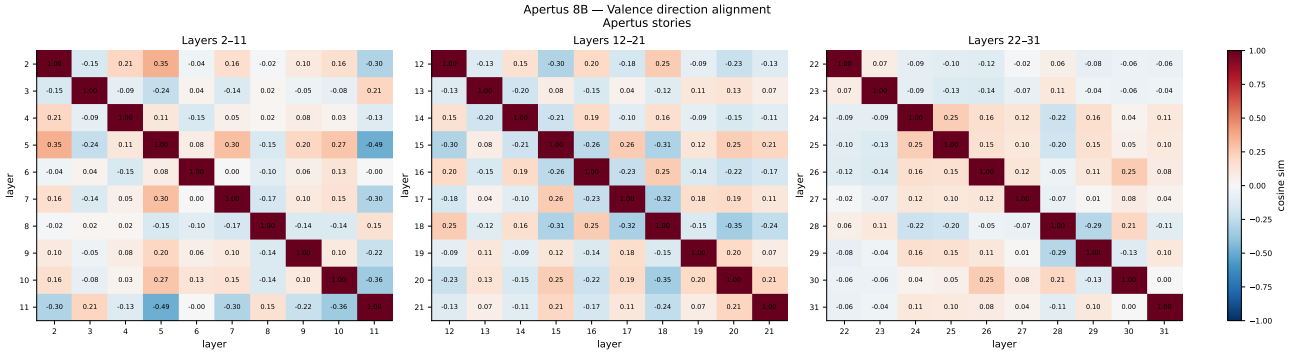


Figure 7. **APERTUS-8B** validation on Gemma stories

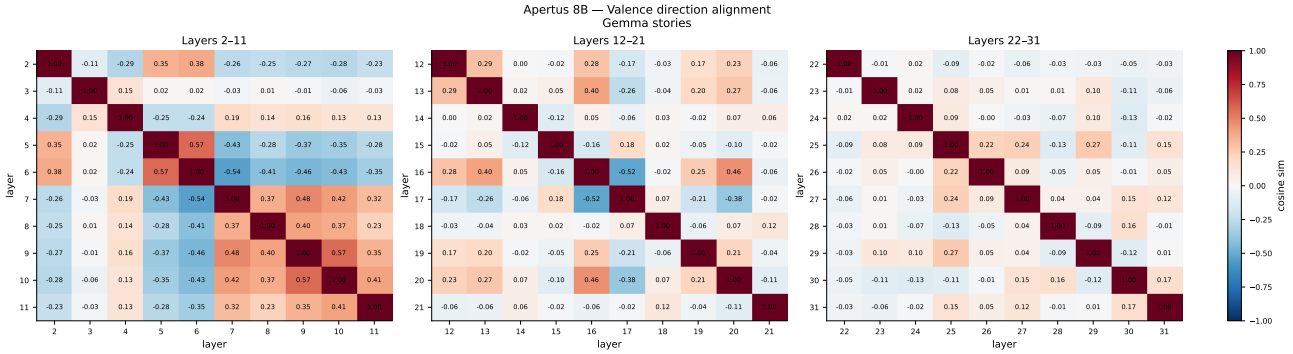


Figure 8. **APERTUS-8B** validation on Apertus stories

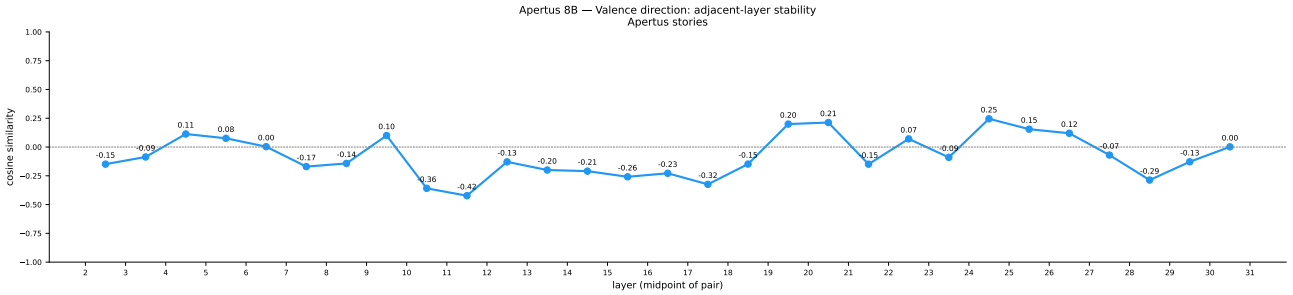
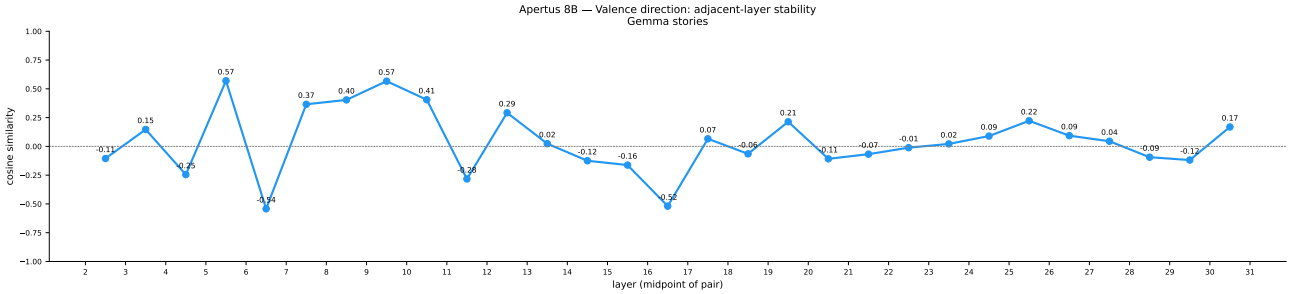


Figure 9. **APERTUS-8B** validation on Gemma stories



C.4.2. GEMMA-4-E4B RESULTS

Figure 10. **GEMMA-4-E4B** validation on Gemma stories

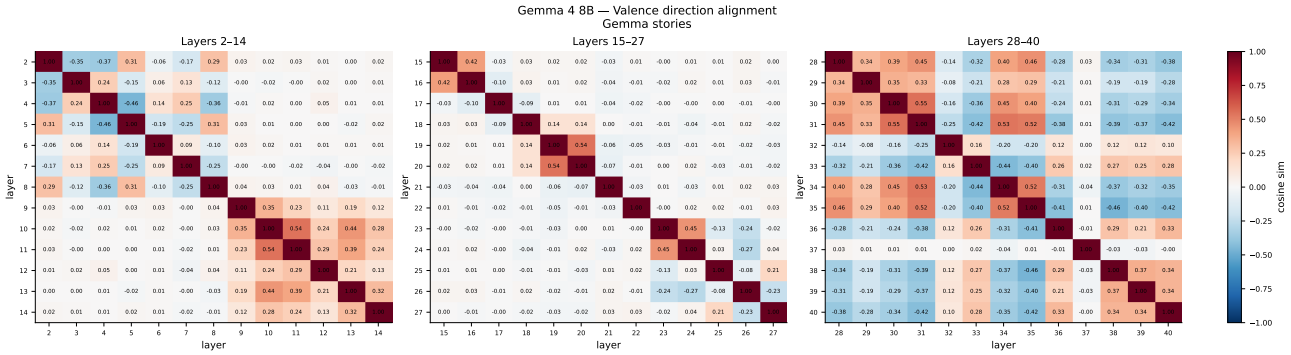


Figure 11. **GEMMA-4-E4B** validation on Apertus stories

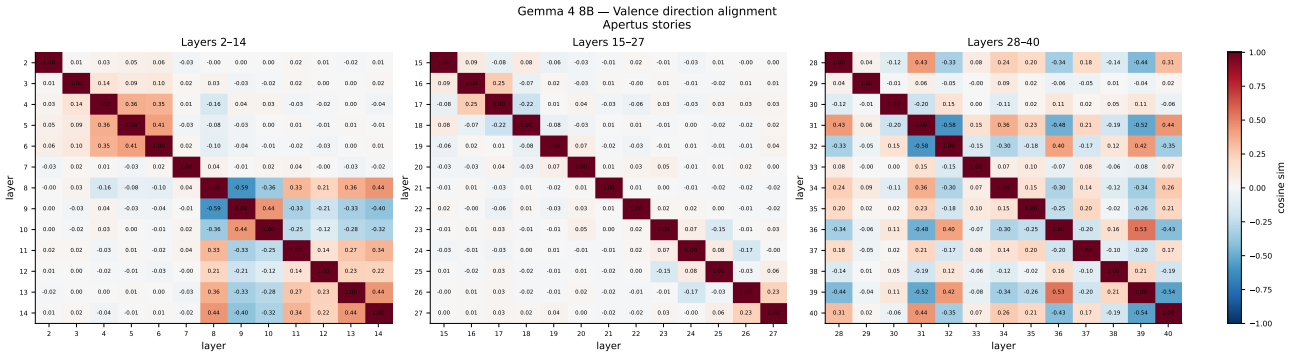


Figure 12. **GEMMA-4-E4B validation** on Apertus stories

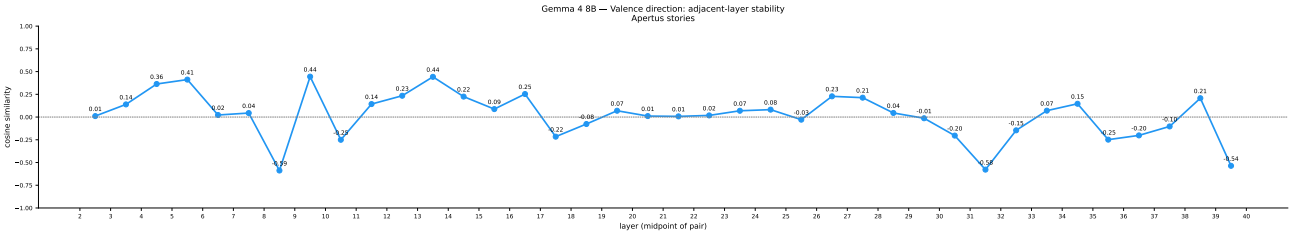
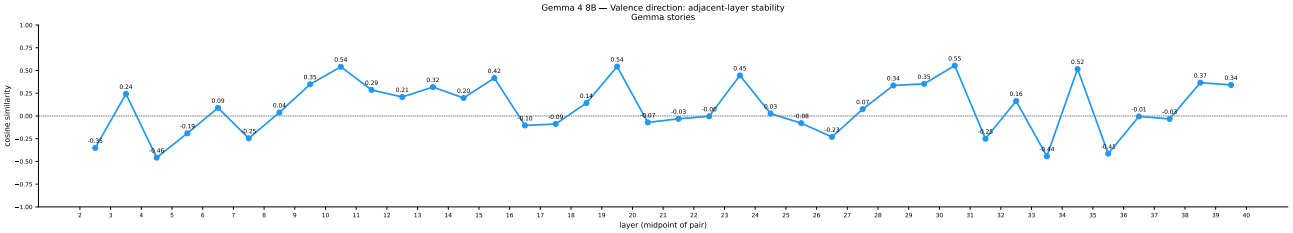


Figure 13. **GEMMA-4-E4B** on Gemma stories



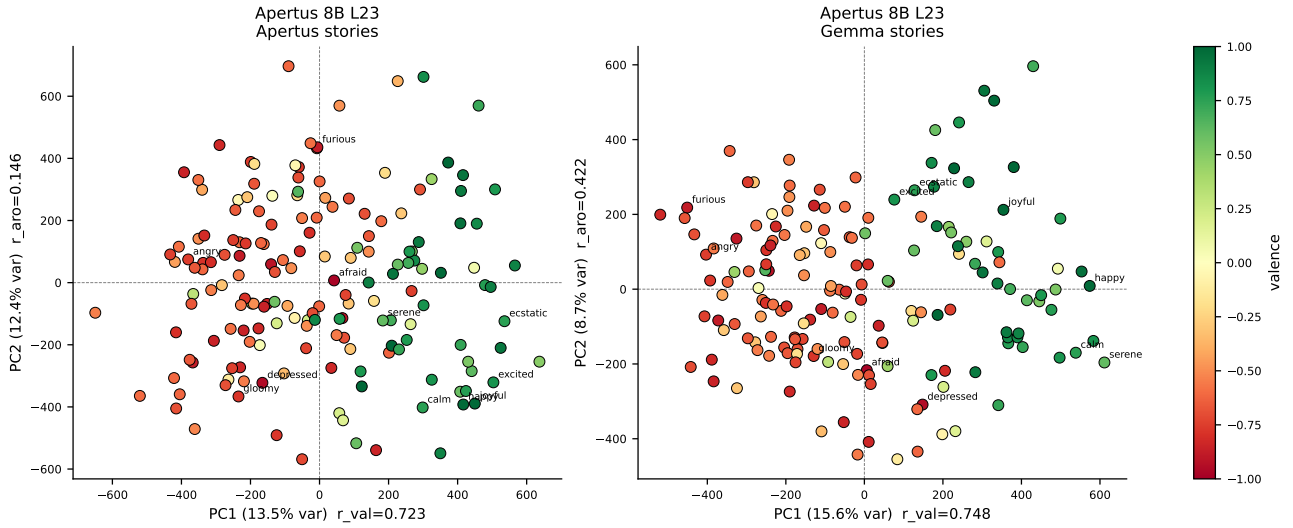
C.5. PCA comparison

The PCA figure shows a map of the model's emotional space at a specific layer. We pick the layer with the highest valence. Each dot is an emotion, positioned at how the model actually represents this emotion in its activation space.

Comparing two panels can tell whether the map is reproducible across different inputs, or whether the emotional space is sensitive to what stories the model reads.

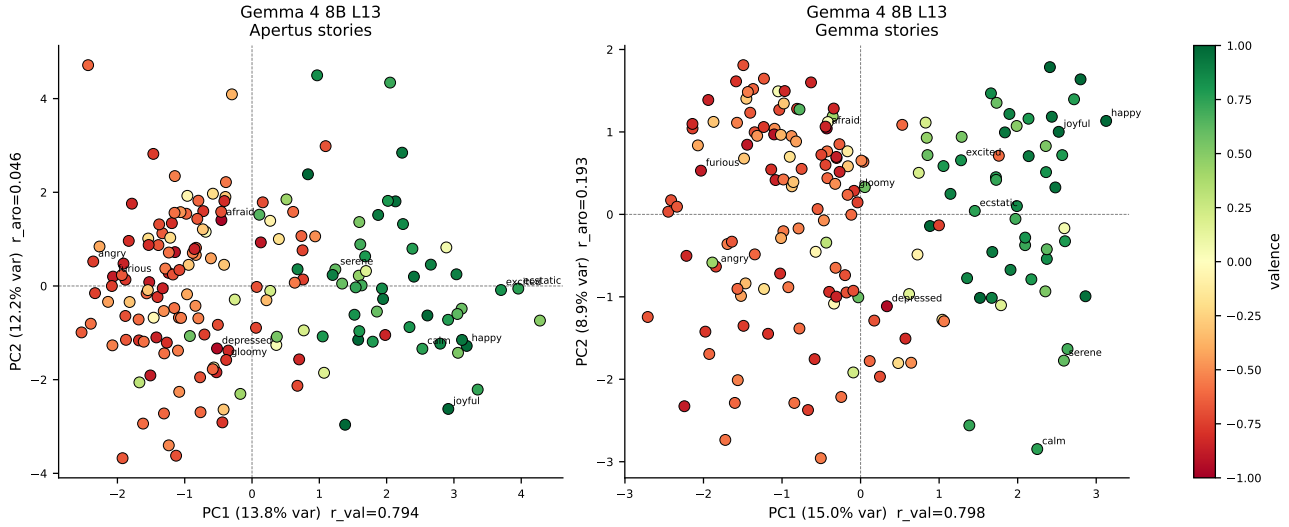
C.5.1. APERTUS-8B RESULTS

Figure 14. **APERTUS-8B** validation



C.5.2. GEMMA-4-E4B RESULTS

Figure 15. **GEMMA-4-E4B: valence-arousal PCA**



D. Prompts

Below, we report verbatim the prompts used to generate the short stories.

System prompt — Explanation generation

```
Write {n_stories} different stories based on the following premise.
Topic: {topic}
The story should follow a character who is feeling {emotion}.
Format the stories like so:
[story 1]
[story 2]
[story 3]
etc.
The paragraphs should each be a fresh start, with no continuity. Try to make them
    diverse and not use the same turns of phrase. Across the different stories, use a
    mix of third-person narration and first-person narration.
IMPORTANT: You must NEVER use the word '{emotion}' or any direct synonyms of it in the
    stories. Instead, convey the emotion ONLY through:
- The character's actions and behaviors
- Physical sensations and body language
- Dialogue and tone of voice
- Thoughts and internal reactions
- Situational context and environmental descriptions
The emotion should be clearly conveyed to the reader through these indirect means, but
    never explicitly named.
```