Emotion Beyond Language: Probing Multilingual Representations in LLMs

Anonymous ACL submission

Abstract

Large language models (LLMs) increasingly process emotional content across languages, but how emotion is internally represented in multilingual LLMs remains unclear. We construct a controlled multilingual narrative set covering 20 emotions in 13 languages, where emotions are expressed through context rather than explicit emotion words. We analyze transformer-layer activations using emotion vectors and layerwise linear probes. Our results show that emotion representations exhibit strong language-specific offsets, yet preserve partially shared relational geometry. Probing further reveals that emotion information is most accessible in middle layers and supports uneven cross-lingual transfer, shaped by language family and resource conditions. These findings provide a framework for studying emotion as an internal multilingual representation.

1 Introduction

Are emotions universal, or shaped by the language and culture that express them? Psychology and linguistics have long disagreed: one tradition treats emotions as biologically basic categories recognized across cultures (Ekman, 1992), organized along a few affective dimensions such as valence and arousal (Russell, 1980; Osgood et al., 1975); another holds that emotions are constructed (Barrett, 2017) and that words like *anger* or *fear* lack exact cross-lingual equivalents (Wierzbicka, 1999; Russell, 1991). Across 2,474 languages, both appear true: emotion semantics vary with cultural and geographic distance, yet all language families organize them primarily by valence and arousal (Jackson et al., 2019).

Multilingual large language models (LLMs) now process emotional content in many languages, learning affective meaning from text with no explicit cross-lingual supervision. This makes them a computational testbed for a representational version of the same question: does an LLM encode each language in a separate emotion space, a shared affective structure, or something in between? The answer matters in practice, as fragmented representations would leave low-resource languages underserved. Yet most emotion work tests only label prediction (Öhman et al., 2020; Muhammad et al., 2025a), while interpretability studies show that LLMs hold emotion-concept representations that generalize and causally shape behavior (Sofroniew et al., 2026). We do not claim that LLMs experience emotions; we study emotion as a functional concept encoded in activations.

We build a controlled multilingual stimulus set of 20 emotions in 13 languages, expressed through context rather than explicit emotion words, and extract decoder-layer activations from six instruction-tuned LLMs. Our contributions are threefold. (i) We introduce a multilingual framework that studies emotion as an internal representation, combining category-level emotion vectors with cross-lingual linear probing. (ii) Emotion vectors are not directly aligned across languages: they carry language-specific offsets in absolute space yet preserve a partially shared relational geometry. (iii) Instance-level hidden states transfer across languages under probing, but unevenly, shaped by language family and resource conditions. Figure 1 overviews our framework.

2 Methodology

2.1 Multilingual Setup and Activation Extraction

We construct a controlled multilingual set of short narratives where emotions are expressed through situations, thoughts, and behavior rather than explicit emotion words. Following the synthetic narrative setup of Sofroniew et al. (2026), we use 100 topics and 20 emotions, extending them to 13 languages: Arabic, Chinese, Dutch, English, French, German, Italian, Japanese, Korean, Por-

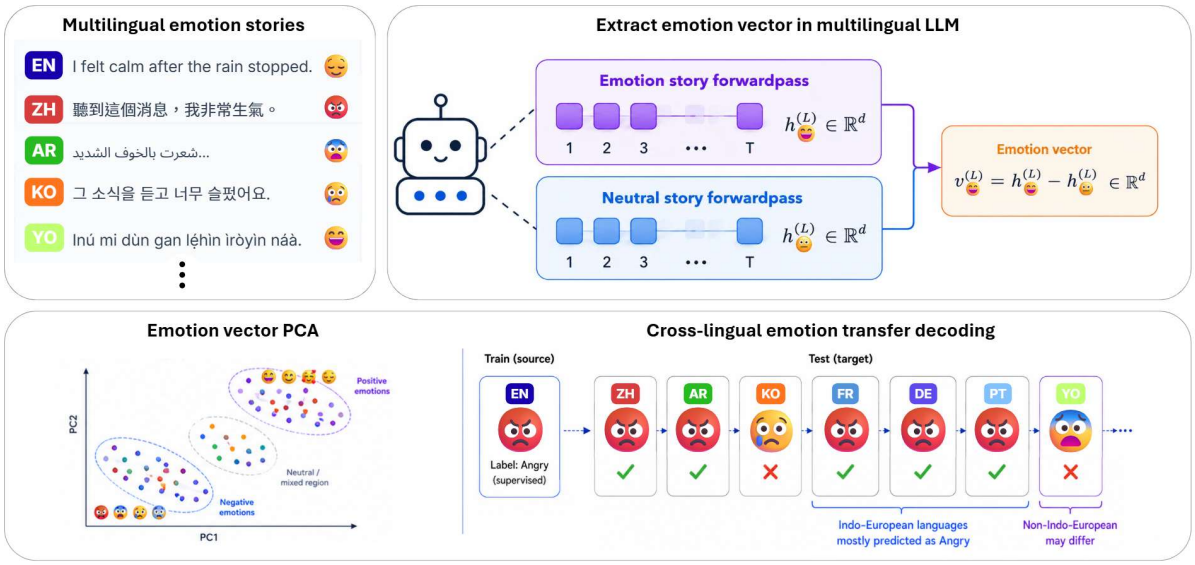


Figure 1: **Framework overview.** Multilingual emotion narratives are forward-passed through the LLM to extract decoder-layer activations, which are analyzed via emotion vectors (left) and cross-lingual linear probes (right), where a probe trained on English with label *Angry* is applied to all other languages. Checkmarks and crosses indicate whether each test language's prediction matches the training label.

tuguese, Spanish, Swahili, and Yoruba. For each language-emotion-topic combination, Gemma-4-31B-IT generates five stories, yielding 130,000 stories. Prompt-facing emotion words are translated for generation, while analyses use canonical English emotion labels. Full prompts, decoding settings, language counts, and data splits are provided in Appendix A.

We evaluate six instruction-tuned LLMs from three model families: Qwen-2.5-3B and 7B (Qwen et al., 2025), Ministral-3-8B and 14B (Liu et al., 2026), and Gemma-4-E2B and E4B. For each story x, we run a forward pass through the target model and extract hidden states from all transformer layers. We remove prompt scaffold tokens and mean-pool the remaining story tokens. The resulting story-level activation at layer l is denoted by $a_l(x)$.

2.2 Emotion Vectors

For each language, emotion, and layer, we estimate an emotion vector from the pooled story activations. Let $\mu_{e,l}$ be the mean activation for emotion e at layer l, and let $\mu_{\text{neu},l}$ be the corresponding neutral mean in the same language. We define a raw emotion direction by subtracting the neutral baseline, then remove leading neutral principal components and L_2-normalize the result:

$$\hat{v}_{e,l} = \text{norm}\left(\text{denoise}\left(\mu_{e,l} - \mu_{\text{neu},l}\right)\right).$$

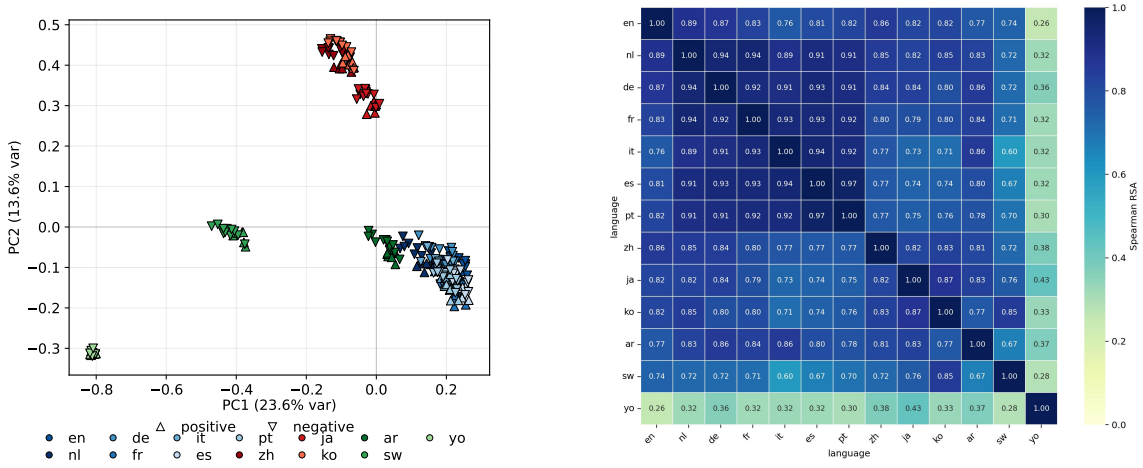
These vectors are used to analyze category-level emotion geometry through joint PCA and representational similarity analysis. The full estimation and denoising procedure is given in Appendix C.

2.3 Supervised Cross-Lingual Linear Probing

Emotion vectors summarize each emotion as a category-level direction. To test whether individual story activations also encode emotion, we train layerwise linear probes on frozen activations. For each training language and transformer layer, we train a multinomial logistic-regression classifier on the topic-disjoint split and evaluate it on held-out activations from all test languages. Diagonal train–test pairs measure within-language decodability, and off-diagonal pairs measure cross-language transfer. We report macro F1 to weight emotion classes equally. Repeating this at every layer yields curves showing where cross-lingual emotion information is most accessible. For transfer-matrix summaries, each language pair is represented by its best validation macro F1 over layers and the corresponding layer index. Implementation details, preprocessing, and hyperparameters appear in Appendix D.

3 Results

Due to space constraints, we report results for Ministral 14B as a representative model. Complete results across all six models and languages are provided in Appendix E.



(a) Joint PCA

(b) Cross-language RSA

Figure 2: Language offsets and relational emotion geometry in Ministral 14B. **Left**: joint PCA of emotion vectors across 13 languages. Colors indicate languages, and markers indicate emotion valence. **Right**: cross-language RSA (Spearman) over pairwise emotion distances. Higher RSA indicates more similar relative emotion geometry.

3.1 Language identity shapes but does not erase emotion geometry

Joint PCA over all 13 languages (Figure 2a) shows that emotion vectors cluster primarily by language rather than by emotion category, indicating strong language-specific offsets in absolute activation space. We therefore ask whether the relative organization of emotions is more stable. Cross-language RSA (Figure 2b) shows that pairwise emotion distances are substantially preserved across languages. RSA ranges from 0.74-1.00 for Indo-European pairs but drops to 0.26-0.38 for pairs involving Yoruba. Multilingual emotion representations thus preserve a partially shared relational geometry under language-specific offsets, with the degree of sharing graded by typological distance.

3.2 Emotion information is cross-lingually decodable in middle layers

3.3 Cross-model Probing Summary

Table 1 aggregates probing performance across all six models. Within-language F1 is consistently high (0.86-0.93) across model families, while cross-lingual transfer is substantially lower (0.36-0.58) and more variable (IQR up to 0.62). Ministral models show the smallest within-to-cross gap, while Gemma models achieve lower cross-lingual median F1 but tighter IQR. Scale effects are modest within each family.

Model	Within-lang median (IQR)	Cross-lang median (IQR)	Gap
Qwen 2.5 3B	0.919 (0.102)	0.462 (0.557)	0.457
Qwen 2.5 7B	0.929 (0.130)	0.512 (0.621)	0.417
Ministral 3 8B	0.919 (0.089)	0.572 (0.327)	0.347
Ministral 3 14B	0.925 (0.095)	0.575 (0.311)	0.351
Gemma 4 E2B	0.863 (0.114)	0.363 (0.213)	0.501
Gemma 4 E4B	0.877 (0.110)	0.399 (0.208)	0.479

Table 1: Cross-model summary of cross-lingual linear probing. Scores are macro F1 aggregated by median. Within-lang covers diagonal train-test pairs, cross-lang covers off-diagonal pairs. Parenthesized values are the IQR of per-pair scores. Probe trained on a single language and evaluated on all 13.

3.4 Language-mean centering reduces transfer mismatch

We further test whether cross-lingual probing is limited by language-specific activation offsets. To do so, we subtract the mean activation of each language from its story representations before training and evaluation, while keeping the same topic-disjoint probing setup. As shown in Figure 4, Chinese-trained probes become more consistently transferable after language-mean centering. Most Indo-European languages, Arabic, Japanese, and Korean form a shared middle-layer plateau, whereas Chinese remains the strongest in-language condition. However, Yoruba stays near chance and Swahili remains weaker, suggesting that language offsets explain part of the transfer gap but not all of it.

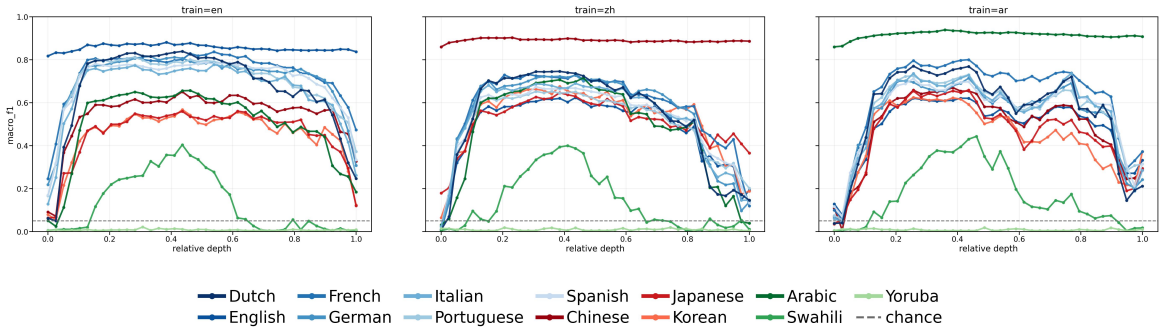


Figure 3: Layerwise cross-lingual emotion probe transfer in Ministral 14B. Each panel uses one training language. Each line shows the macro-F1 of the probe on one test language. The dashed line indicates chance performance for the 20-way emotion classification task.

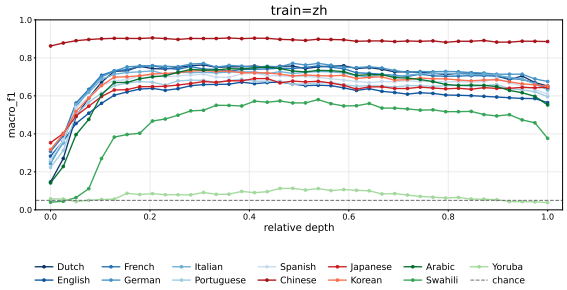


Figure 4: Chinese-trained cross-lingual probe transfer after language-mean centering. Most languages align in middle layers, while Yoruba and Swahili remain weak.

4 Related Work

Multilingual Emotion Resources. Text-based emotion recognition has expanded from fine-grained English datasets (Demszky et al., 2020) to broad multilingual benchmarks (Öhman et al., 2020; Muhammad et al., 2025a,b), but these efforts evaluate label prediction from model outputs. We instead examine how emotion is encoded in hidden states and whether it transfers across languages.

Emotion representations inside LLMs. Affective information is linearly encoded in neural representations, from an early sentiment neuron (Radford et al., 2017) to decodable sentiment and emotion in intermediate hidden states (Tigges et al., 2023; Di Palma et al., 2025), and representation engineering shows that emotion-like concepts form separable, manipulable directions (Zou et al., 2023; Marks and Tegmark, 2023). Most closely related to us, Sofroniew et al. (2026) identify internal emotion concepts in an LLM and demonstrate their causal influence on behavior. Whether such representations are language-specific or shared across languages remains open.

Multilingual representation analysis. Probing is a standard tool for analyzing neural representations (Belinkov, 2022), and multilingual models support cross-lingual transfer despite imperfect alignment (Pires et al., 2019; Wu and Dredze, 2019), mixing language-neutral and language-specific components depending on linguistic distance and the information probed (Libovický et al., 2019; Conneau et al., 2020). Geometry-based analysis has targeted syntax and semantics (Chi et al., 2020), but affective meaning less so. We combine probing with vector-level geometry to ask whether emotion is partially shared across languages or carries strong language-specific offsets.

5 Conclusion

We presented a controlled multilingual analysis of emotion representations in instruction-tuned LLMs. Across 20 emotions and 13 languages, emotion representations are neither fully language-independent nor entirely language-specific. At the category level, emotion vectors show strong language-specific offsets in absolute activation space while preserving partially shared relational geometry. At the instance level, layerwise linear probes show that emotion information is most accessible in middle layers and transfers across languages, though transfer strength varies with language family and resource conditions. These findings suggest that multilingual LLMs encode affective meaning in a structured but imperfectly aligned representational space. More broadly, our work offers a framework for studying emotion as an internal multilingual representation rather than only an output-level classification problem.

Limitations

Data and model scope. Our multilingual stimulus set is generated by a single model (Gemma 4 31B IT), which may introduce systematic biases in how emotions are lexicalized and expressed, particularly for low-resource languages such as Yoruba and Swahili where the generator may rely on code-switching rather than authentic in-language conventions. Furthermore, our analysis is restricted to instruction-tuned models; whether the observed representational structure generalizes to base models or models trained under different alignment procedures remains an open question, as does whether category-level emotion vectors—derived from synthetic narratives—reflect the full diversity of emotional contexts present in pretraining data. Methodological limitations. Cross-lingual linear probing establishes that emotion information is decodable across languages but does not demonstrate that the shared representational structure is causally implicated in model behavior; establishing such claims would require combining activation steering with multilingual behavioral evaluation. Additionally, our probe evaluations report macro-F1 averaged uniformly over all 20 emotion classes, which obscures variation among semantically proximate pairs (e.g., sad vs. gloomy, angry vs. hostile) that may be structurally harder to transfer; fine-grained per-emotion transfer analysis is left for future work.

Use of AI Assistants

We used ChatGPT to help refine phrasing, improve clarity, and enhance the structure and readability of the manuscript. The content remains entirely our own.

References

Lisa Feldman Barrett. 2017. The theory of constructed emotion: an active inference account of interoception and categorization. *Social Cognitive and Affective Neuroscience*, 12:1833 – 1833.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Ethan A Chi, John Hewitt, and Christopher D Manning. 2020. Finding universal grammatical relations in multilingual bert. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5564–5577.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6022–6034, Online. Association for Computational Linguistics.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 4040–4054.

Dario Di Palma, Alessandro De Bellis, Giovanni Servedio, Vito Walter Anelli, Fedelucio Narducci, and Tommaso Di Noia. 2025. Llamas have feelings too: Unveiling sentiment and emotion representations in llama models through probing. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6124–6142.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & Emotion*, 6:169–200.

Joshua Conrad Jackson, Joseph Watts, Teague Rhine Henry, Johann-Mattis List, Robert Forkel, Peter J. Mucha, Simon J. Greenhill, Russell David Gray, and Kristen A. Lindquist. 2019. Emotion semantics show both cultural variation and universal structure. *Science*, 366:1517 – 1522.

Jindřich Libovický, Rudolf Rosa, and Alexander Fraser. 2019. How language-neutral is multilingual bert? *Preprint*, arXiv:1911.03310.

Alexander H. Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, Adrien Sadé, Alan Jeffares, Albert Jiang, Alexandre Cahill, Alexandre Gavaudan, Alexandre Sablayrolles, Amélie Héliou, Amos You, Andy Ehrenberg, Andy Lo, Anton Eliseev, Antonia Calvi, Avinash Sooriyarachchi, Baptiste Bout, and 101 others. 2026. Ministral 3. *Preprint*, arXiv:2601.08584.

Samuel Marks and Max Tegmark. 2023. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:2310.06824*.

Shamsuddeen Hassan Muhammad, Nedjma Ousidhoum, Idris Abdulmumin, Jan Philip Wahle, Terry Ruas, Meriem Beloucif, Christine De Kock, Nirmal Surange, Daniela Teodorescu, Ibrahim Sa'id Ahmad, and 1 others. 2025a. Brighter: Bridging the gap in human-annotated textual emotion recognition datasets for 28 languages. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8895–8916.

Shamsuddeen Hassan Muhammad, Nedjma Ousidhoum, Idris Abdulmumin, Seid Muhie Yimam, Jan Philip Wahle, Terry Lima Ruas, Meriem Beloucif, Christine De Kock, Tadesse Destaw Belay, Ibrahim Sa'id Ahmad, and 1 others. 2025b. Semeval-2025 task 11:

Bridging the gap in text-based emotion detection. In *Proceedings of the 19th international workshop on semantic evaluation (SemEval-2025)*, pages 2558–2569.

Emily Öhman, Marc Pàmies, Kaisla Kajava, and Jörg Tiedemann. 2020. Xed: A multilingual dataset for sentiment analysis and emotion detection. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6542–6552.

Charles E. Osgood, William H. May, and Murray S. Miron. 1975. Cross-cultural universals of affective meaning.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 4996–5001.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Alec Radford, Rafal Jozefowicz, and Ilya Sutskever. 2017. Learning to generate reviews and discovering sentiment. *Preprint*, arXiv:1704.01444.

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

James A. Russell. 1991. Culture and the categorization of emotions. *Psychological bulletin*, 110 3:426–50.

Nicholas Sofroniew, Isaac Kauvar, William Saunders, Runjin Chen, Tom Henighan, Sasha Hydrie, Craig Citro, Adam Pearce, Julius Tarng, Wes Gurnee, Joshua Batson, Sam Zimmerman, Kelley Rivoire, Kyle Fish, Chris Olah, and Jack Lindsey. 2026. Emotion concepts and their function in a large language model. *Preprint*, arXiv:2604.07729.

Curt Tigges, Oskar John Hollinsworth, Atticus Geiger, and Neel Nanda. 2023. Linear representations of sentiment in large language models. *Preprint*, arXiv:2310.15154.

Anna Maria Wierzbicka. 1999. Emotions across languages and cultures: Diversity and universals: Defining emotion concepts: discovering "cognitive scenarios".

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of bert. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 833–844.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, and 1 others. 2023. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*.

A Details of Multilingual Stimulus Generation

Index	Emotion	Valence	Arousal
1	happy	positive	0.65
2	blissful	positive	0.45
3	proud	positive	0.60
4	hopeful	positive	0.45
5	grateful	positive	0.35
6	loving	positive	0.45
7	calm	positive	0.10
8	enthusiastic	positive	0.90
9	sad	negative	0.20
10	gloomy	negative	0.25
11	brooding	negative	0.35
12	reflective	negative	0.20
13	guilty	negative	0.35
14	nervous	negative	0.70
15	anxious	negative	0.80
16	afraid	negative	0.90
17	angry	negative	0.90
18	hostile	negative	0.85
19	exasperated	negative	0.75
20	desperate	negative	0.95

Table 2: The 20 emotion categories used for stimulus generation and analysis, with assigned valence labels and approximate arousal levels based on Russell's circumplex model of affect.

A.1 Languages and Corpus Size

Table 3 summarizes the languages used for stimulus generation. Each language contains 20 emotions (Table 2, 100 topics, and 5 stories for each emotion-topic pair. For each language and layer, the neutral baseline $\mu_{\text{neu},l}$ is computed from a matched set of neutral stimulus texts. These texts are generated using the same model and decoding settings as the emotional narratives, covering the same 100 topics, but without specifying any target emotion. Each topic yields five neutral texts per language, resulting in 500 neutral texts per language in total. By matching topics across emotional and neutral conditions, the difference-in-means direction captures emotional content rather than topic- or domain-level variation.

A.2 Translation of Topics and Emotion Words

We translate both the 100 topics and the prompt-facing emotion words into each target language. The translated emotion words are used to instruct

Language	Code	Number of stories
Arabic	ar	10,000
Chinese	zh	10,000
Dutch	nl	10,000
English	en	10,000
French	fr	10,000
German	de	10,000
Italian	it	10,000
Japanese	ja	10,000
Korean	ko	10,000
Portuguese	pt	10,000
Spanish	es	10,000
Swahili	sw	10,000
Yoruba	yo	10,000
Total		130,000

Table 3: Languages and corpus size for multilingual emotion stimuli.

the generator, while the analysis labels remain canonical English labels. For example, the English emotion label *happy* may be translated into Chinese in the generation prompt, but the resulting story is still stored under the canonical label *happy*.

A.3 Generation Prompt

For each language, emotion, and topic, we use a prompt template that asks Gemma 4 31B to write a short story in the target language. The prompt specifies the target emotion and topic, and instructs the model not to mention the target emotion word directly.

```
Write a short paragraph (4-6
sentences) about topic. In the
story, the main character is
clearly feeling emotio_word.
Show the emotion through their
internal thoughts, physical
sensations, or behavior. Do
NOT use the word "emotion_word"
anywhere in the text. Do NOT
start with "The character" or
use generic openings. Return
only the story paragraph; do not
include analysis, planning, or
thinking tags.
```

A.4 Generation Parameters

Table 4 reports the decoding and API settings used for story generation.

Parameter	Value
Generator model	Gemma 4 31B `google/gemma-4-31b-it`
Temperature	0.8
Top-p	0.9
Maximum tokens	2048
Filtering	Strip `<think>`/reasoning blocks

Table 4: Generation settings for multilingual emotion stimuli.

A.5 Data Split for Probing

For supervised probing experiments, we split the 100 topics into 80 training topics and 20 validation topics. The split is applied by topic rather than by story, preventing the probe from seeing the same topic during both training and evaluation. The split is applied independently within each language and emotion, and all validation results are computed on the held-out topic split.

B Model and Activation Extraction Details

We evaluate six instruction-tuned models from three model families: Qwen 2.5 3B, Qwen 2.5 7B, Ministral 3 8B, Ministral 3 14B, Gemma 4 E2B, and Gemma 4 E4B. All analyses freeze the target model parameters and use forward-pass activations only.

For each generated story x, we extract hidden states from all transformer layers. We discard prompt scaffold tokens before pooling, so the representation reflects the story content rather than the instruction template. The remaining story-token hidden states are mean-pooled to obtain a single story-level activation $a_l(x)$ at each layer l. Layer-wise probing is performed independently for every transformer layer; for visualization, layer indices are also reported as relative depth.

C Emotion Vector Estimation Details

This section gives the full definition of emotion-vector estimation. For each language and layer l, we first compute the mean activation of stories labeled with emotion e,

$$\mu_{e,l} = \frac{1}{N_e} \sum_{i=1}^{N_e} a_l(x_{e,i}).$$

We then subtract the mean activation of neutral prompts from the same language and layer,

$$v_{e,l}^{\text{raw}} = \mu_{e,l} - \mu_{\text{neu},l}.$$

To reduce variation that is not specific to emotion, we remove dominant directions estimated

from neutral prompts. For each language and layer l, we collect the neutral story activations and compute their principal components. Let $U_l \in \mathbb{R}^{d \times k}$ denote the matrix containing the top k principal components, where k is chosen as the smallest number of components that explain at least 50% of the variance in neutral activations. Given $v_{e,l}^{\text{raw}}$, we remove its projection onto the neutral principal subspace,

$$\tilde{v}_{e,l} = \left(I - U_l U_l^\top\right) v_{e,l}^{\text{raw}}.$$

The final emotion vector is obtained by normalizing the denoised direction,

$$\hat{v}_{e,l} = \frac{\tilde{v}_{e,l}}{\|\tilde{v}_{e,l}\|_2}.$$

This step removes directions that are common among neutral prompts, which may reflect topic, syntax, sequence length, or prompt-format variation rather than emotion-specific information.

D Linear Probe Implementation Details

We implement the linear probes using multinomial logistic regression with L2 regularization, fitted with scikit-learn's `LogisticRegression` using the `lbfgs` solver and the multinomial objective.[1] The regularization strength is set to $C = 1.0$, `max_iter` $= 2000$, `random_state` $= 42$, and no class weighting is applied (`class_weight=None`).

For preprocessing, we use per-feature standardization (`StandardScaler`, i.e. `-feature_norm` $=$ `standardize` in our pipeline).

Crucially, the `StandardScaler` statistics (per-feature means and scales) are fitted only on the training-language training activations, separately for each (model, train language, layer) condition, and then applied unchanged to all test-language activations.

We train a single probe per condition with a fixed seed (`seed=42`); results are not averaged over multiple random seeds, since the dominant source of variation in our experimental design is the (train-language, test-language, layer) triple rather than logistic-regression initialization, and the solver is convex.

For each (train language, test language, layer) we compute macro F1 on the validation split. The reported transfer matrix uses the best validation macro F1 over layers for each language pair (ties broken by accuracy, then by smallest layer index). We also record the corresponding best layer.

[1] Implemented in `src/probing.py::train_linear_probe`. For scikit-learn versions that still expose `multi_class`, we set it explicitly to `"multinomial"`; for scikit-learn ≥ 1.8, where `multi_class` was removed, `lbfgs` uses the multinomial objective for multiclass problems by default.

E More Detailed Results

E.1 Additional observations on cross-language geometry

Figures 5- 10 present per-language within-language PCA projections across all 13 languages for each of the six evaluated models (Ministral 3 14B, Ministral 3 8 , Gemma 4 E4B, Gemma 4 E2B, Qwen 2.5 7B, Qwen 2.5 3B). Unlike the joint PCA in Figure 2a, which projects all languages into a single shared coordinate system and therefore reveals language-level offsets, here we fit PCA independently for each language. This removes the dominant language-specific offset and exposes the local relational organization of emotion vectors within each language. Several observations are consistent across models. First, the first principal component broadly separates emotions by valence in the majority of languages: positive emotions (e.g., happy, blissful, hopeful) tend to cluster on one side of the projection, while negative emotions (e.g., afraid, angry, desperate) cluster on the opposite side. This replicates the valence-dominant structure observed in English and confirms that it is not an English-specific artifact but a general property of how these models organize affective meaning. Second, within each model, the within-language geometry is broadly similar across languages, despite the models never seeing language-parallel supervision during training. This supports the interpretation that the shared relational structure observed in RSA reflects genuine cross-lingual commonality rather than a measurement artifact. At the same time, the projections are not identical across languages or models. Languages such as Yoruba and Swahili tend to show more compressed and less differentiated layouts, consistent with their weaker RSA alignment scores and lower probing transfer rates. Across model families, larger models within the same family produce more separated and interpretable within-language geometries, as discussed further in §E.2. Figure 11 provides a focused view of this within-language structure by overlaying the PCA projections of English, Arabic, and Chinese at the best-performing middle layer of Ministral 14B. These three languages are selected as representatives of distinct language families (Germanic,

Figure 5: Per-language within-language PCA of emotion vectors across all 13 languages in **Ministral 14B IT**. Each subplot fits PCA independently on the emotion vectors of a single language. Colors indicate emotion valence (positive / negative). This visualization removes cross-language offsets and reveals the local relational organization of emotions within each language. A consistent valence-like structure emerges across most languages, though the precise layout varies by language.

Figure 6: Per-language within-language PCA of emotion vectors across all 13 languages in **Ministral 8B IT**. Each subplot fits PCA independently on the emotion vectors of a single language. Colors indicate emotion valence (positive / negative). This visualization removes cross-language offsets and reveals the local relational organization of emotions within each language. A consistent valence-like structure emerges across most languages, though the precise layout varies by language.

Figure 7: Per-language within-language PCA of emotion vectors across all 13 languages in **Gemma 4 E4B IT**. Each subplot fits PCA independently on the emotion vectors of a single language. Colors indicate emotion valence (positive / negative). This visualization removes cross-language offsets and reveals the local relational organization of emotions within each language. A consistent valence-like structure emerges across most languages, though the precise layout varies by language.

Figure 8: Per-language within-language PCA of emotion vectors across all 13 languages in **Gemma 4 E2B IT**. Each subplot fits PCA independently on the emotion vectors of a single language. Colors indicate emotion valence (positive / negative). This visualization removes cross-language offsets and reveals the local relational organization of emotions within each language. A consistent valence-like structure emerges across most languages, though the precise layout varies by language.

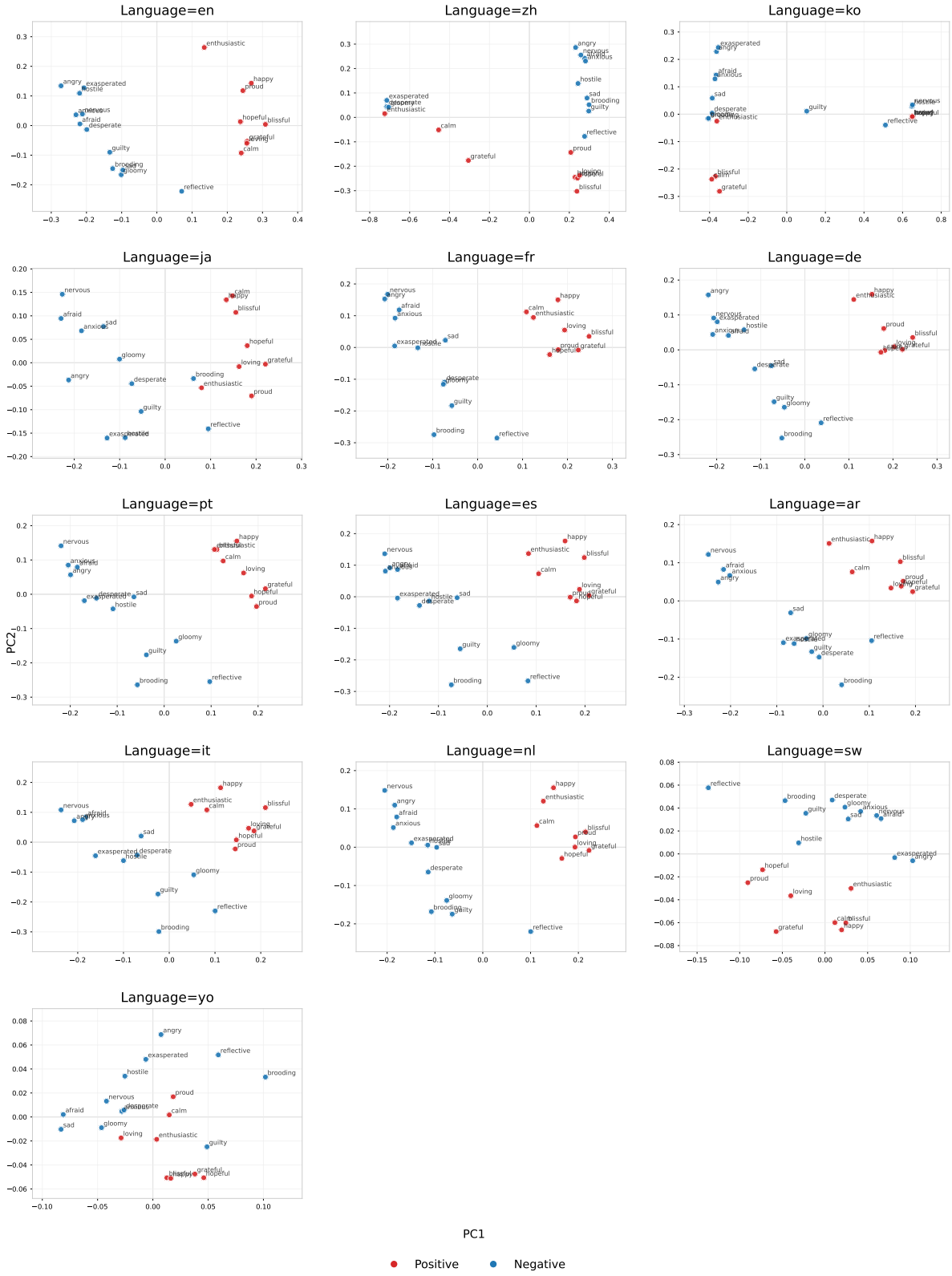


Figure 9: Per-language within-language PCA of emotion vectors across all 13 languages in **Qwen 2.5 7B IT**. Each subplot fits PCA independently on the emotion vectors of a single language. Colors indicate emotion valence (positive / negative). This visualization removes cross-language offsets and reveals the local relational organization of emotions within each language. A consistent valence-like structure emerges across most languages, though the precise layout varies by language.

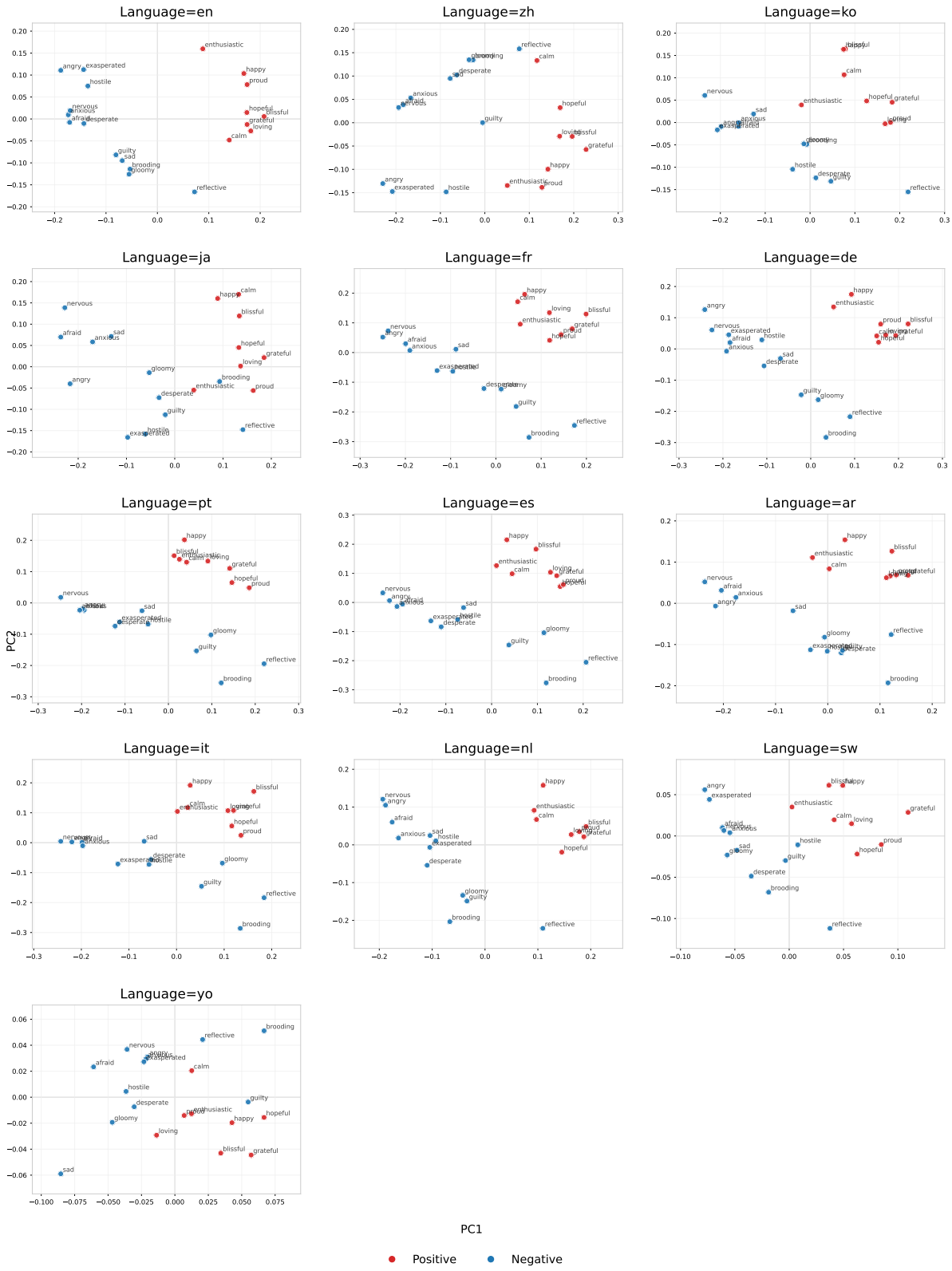


Figure 10: Per-language within-language PCA of emotion vectors across all 13 languages in **Qwen 2.5 3B IT**. Each subplot fits PCA independently on the emotion vectors of a single language. Colors indicate emotion valence (positive / negative). This visualization removes cross-language offsets and reveals the local relational organization of emotions within each language. A consistent valence-like structure emerges across most languages, though the precise layout varies by language.

Afro-Asiatic, and Sino-Tibetan respectively). The overlay confirms that the valence-like first principal component is preserved across all three, while the precise placement of individual emotion categories varies, indicating that the shared geometry is approximate rather than fully language-invariant.

E.2 Emotion geometry is stable within model families

We ask whether the geometry of emotion vectors is preserved across model scale. Figure 12 overlays PCA projections from models within the same family for English. Each model is projected separately, and PCA axes are aligned before overlaying models from the same family.

Across Ministral, Gemma, and Qwen, models from the same family show highly similar emotion layouts. In Ministral, the smaller and larger models place positive emotions such as *happy*, *blissful*, *proud*, and *hopeful* on one side of the space. Negative emotions such as *afraid*, *angry*, *hostile*, and *desperate* appear on the opposite side. Qwen shows a similar cross-scale alignment. Gemma also preserves a clear valence structure, although its geometry is more compressed along a dominant axis.

These results indicate that emotion-vector geometry is not arbitrary at the level of individual checkpoints. The relative organization of emotions is largely preserved within a model family across scale. Larger models also tend to show clearer emotion separation. This suggests that scale strengthens the recoverability of emotion signals while preserving family-specific geometry.

E.3 Probing Curves Across All Models

Figures 13–18 present the full layerwise cross-lingual emotion probe transfer results for the remaining five models (Ministral 3 8B, Qwen 2.5 3B, Qwen 2.5 7B, Gemma 4 E2B, Gemma 4 E4B); results for Ministral 3 14B are shown in the main paper (Figure 3). Each panel corresponds to one training language; each line shows the macro-F1 of the probe evaluated on one test language as a function of relative layer depth. The dashed line marks chance performance for the 20-way classification task (≈ 0.05).

Layerwise profile shape varies across model families. The most salient architectural difference across models lies in the shape of the layerwise transfer curve rather than in the overall level of performance. Ministral 3 8B and 14B both produce a broad, smooth plateau: performance rises from near chance in the earliest layers, plateaus across a wide middle band (approximately 0.3–0.7 relative depth), and declines near the final layers. Qwen 2.5 3B and 7B show a broadly similar bell shape, but the plateau is narrower and the rise from early layers is steeper, with some local irregularities. Gemma 4 E2B and E4B exhibit a qualitatively distinct multi-peak profile: a sharp early rise produces a local peak around 0.15–0.20 relative depth, followed by a dip and a secondary, often higher peak around 0.40–0.55 relative depth, before a final decline. This multi-peak structure is visible across nearly all training languages in both Gemma variants, suggesting it reflects an architectural property of the model rather than noise. One possible source is the mixture-of-experts routing in Gemma 4, which may create discontinuities in information flow across depth; we flag this as a hypothesis rather than a confirmed explanation.

A further distinctive property of Gemma 4 is that within-language performance for several languages (notably Chinese, Korean, Japanese, French, and Arabic) reaches high levels already in the earliest layers—at relative depths where Ministral and Qwen probes remain near chance. This early-layer accessibility of emotion information in Gemma 4 may reflect differences in tokenization, positional encoding, or the effective depth at which contextual integration begins in MoE architectures.

Language-family structure is preserved across all models. Despite the differences in curve shape, the ranking of cross-lingual transfer by language family is consistent across all six models. Indo-European test languages (Dutch, English, French, German, Italian, Portuguese, Spanish) form a coherent high-performing cluster when a probe is trained on any Indo-European language. East Asian languages (Chinese, Japanese, Korean) show stronger mutual transfer among themselves than to languages from other families; this within-cluster cohesion is particularly pronounced in Gemma and in smaller Qwen models, where the gap between East Asian and Indo-European test performance is larger than in Ministral 14B.

Scale effects within families. Within each model family, increasing scale consistently raises both within-language F1 and cross-lingual transfer ceilings. The effect is clearest in Ministral: the 14B model produces a broader plateau and higher cross-

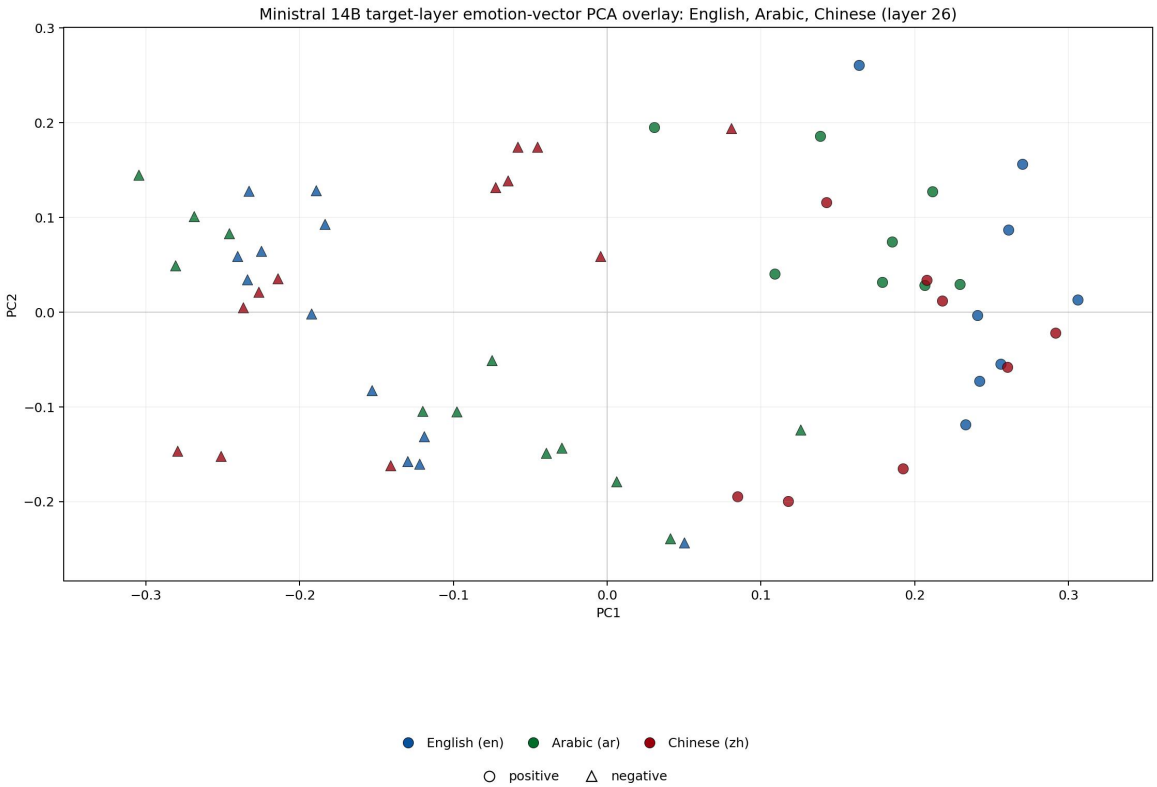


Figure 11: Within-language emotion-vector PCA overlay for English (en), Arabic (ar), and Chinese (zh) at the target layer (layer 24) of Ministral 14B. Each language is projected separately using its own PCA axes; the three projections are then overlaid for comparison. Triangle markers indicate negative-valence emotions; circle markers indicate positive-valence emotions. The first principal component broadly separates positive (right) from negative (left) emotions across all three languages, suggesting a partially shared valence-like structure beneath language-specific variation.

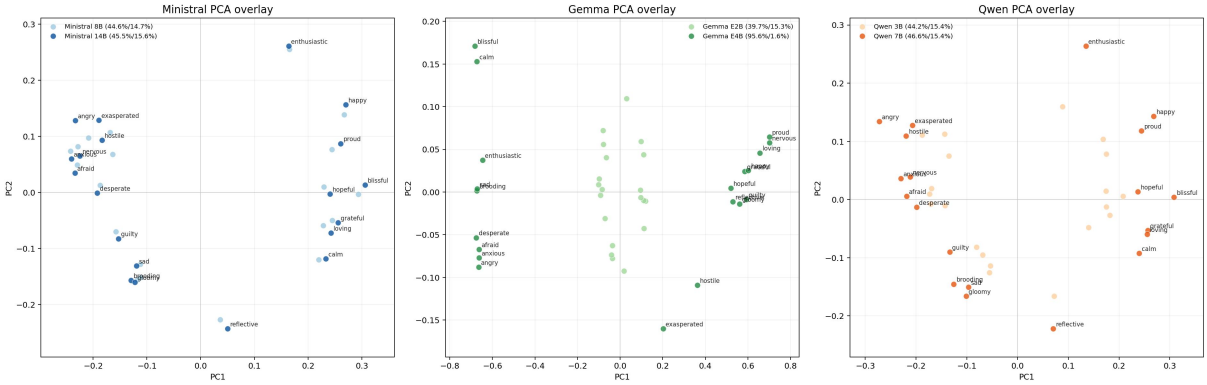


Figure 12: Cross-scale PCA overlay for English emotion vectors. Each panel corresponds to one model family. Within each family, two model scales are projected separately and aligned by PCA axis direction. No cross-model scale normalization is applied.

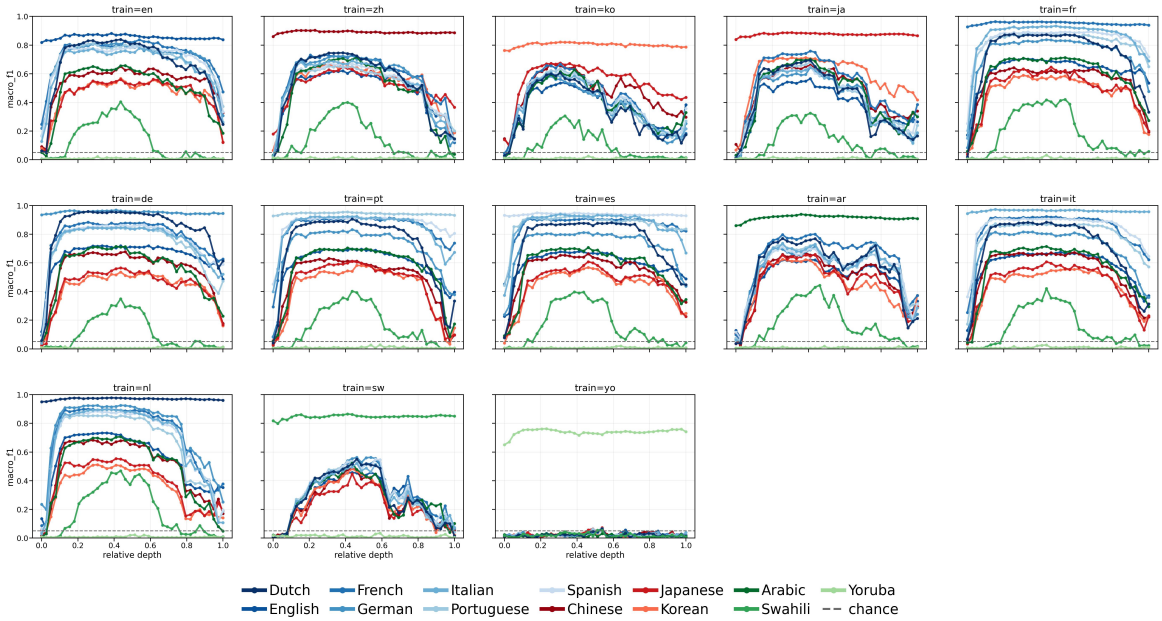


Figure 13: Layerwise cross-lingual emotion probe transfer in **Ministral 3 14B IT**. Each panel uses one training language; each line shows macro-F1 on one test language. The dashed line indicates chance performance (≈ 0.05).

lingual F1 than the 8B model across nearly all language pairs. In Qwen, the 7B model yields meaningfully higher transfer to distant languages (Arabic, East Asian) compared to the 3B model. The improvement from Gemma E2B to E4B is more modest, with the multi-peak shape preserved and only small gains in absolute F1.

Low-resource language asymmetry. Yoruba and Swahili show a consistent asymmetric pattern across all models. Both languages achieve above-chance within-language decodability (Yoruba: 0.60–0.80; Swahili: 0.75–0.85 depending on model), indicating that emotion information is encoded in their activations. However, cross-lingual transfer from Yoruba to all other languages is at or near chance in every model, and transfer from other languages to Yoruba as a test language is likewise negligible. This bidirectional isolation suggests that the emotion representations of Yoruba are encoded in a subspace poorly aligned with those of all other evaluated languages, rather than merely being weaker in magnitude.

Swahili presents a different profile: while cross-lingual transfer from Swahili is generally weak, Arabic consistently achieves substantially above-chance performance as a test language when a probe is trained on Swahili. This Swahili $\rightarrow$ Arabic transfer is visible across all six models and is not reciprocated symmetrically (Arabic-trained probes

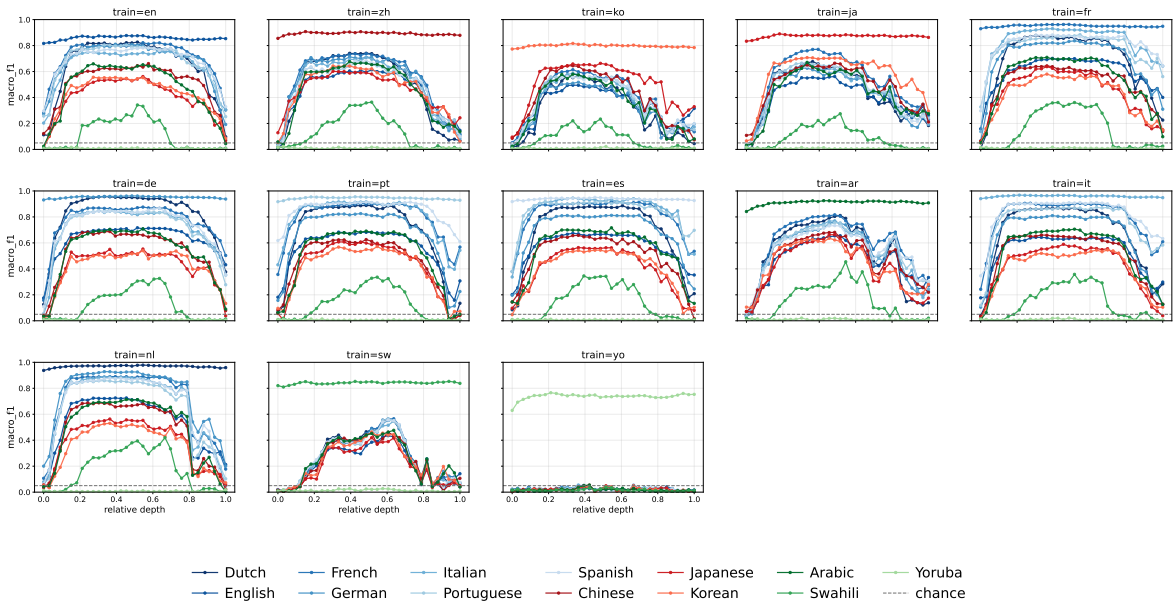


Figure 14: Layerwise cross-lingual emotion probe transfer in **Ministral 3 8B IT**. Each panel uses one training language; each line shows macro-F1 on one test language. The dashed line indicates chance performance (≈ 0.05).

do not show uniformly elevated Swahili transfer). This asymmetry warrants further investigation; one possibility is that Swahili's Arabic-derived lexical layer leaves traces in the model's intermediate representations that are decodable by a probe aligned to Arabic emotion geometry.

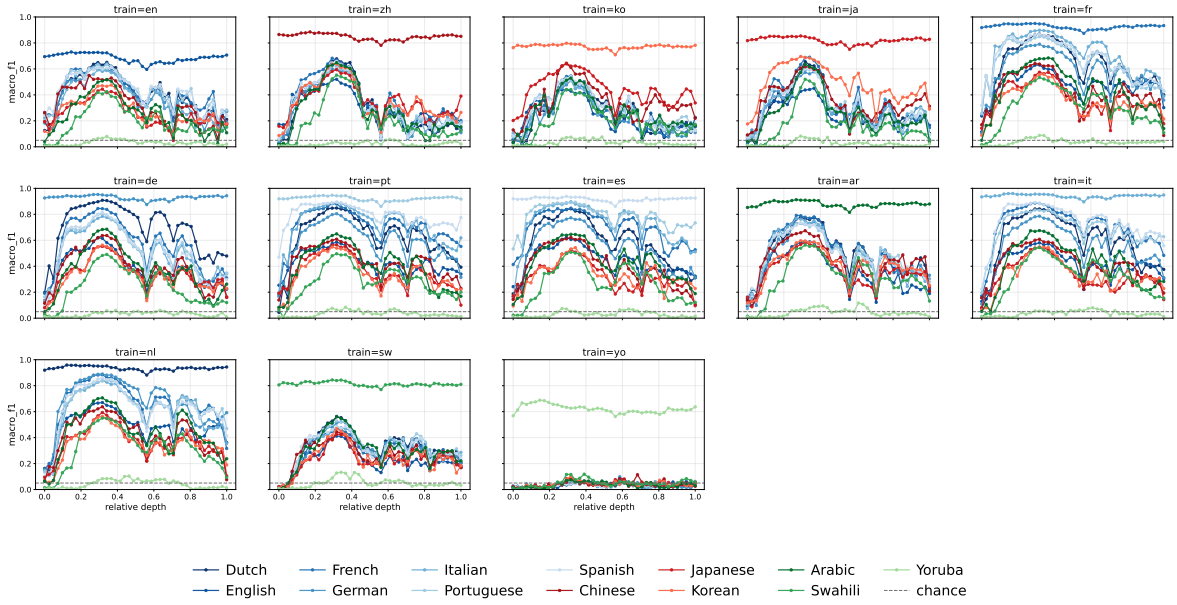


Figure 15: Layerwise cross-lingual emotion probe transfer in **Gemma 4 E4B IT**. Each panel uses one training language; each line shows macro-F1 on one test language. The dashed line indicates chance performance (≈ 0.05).

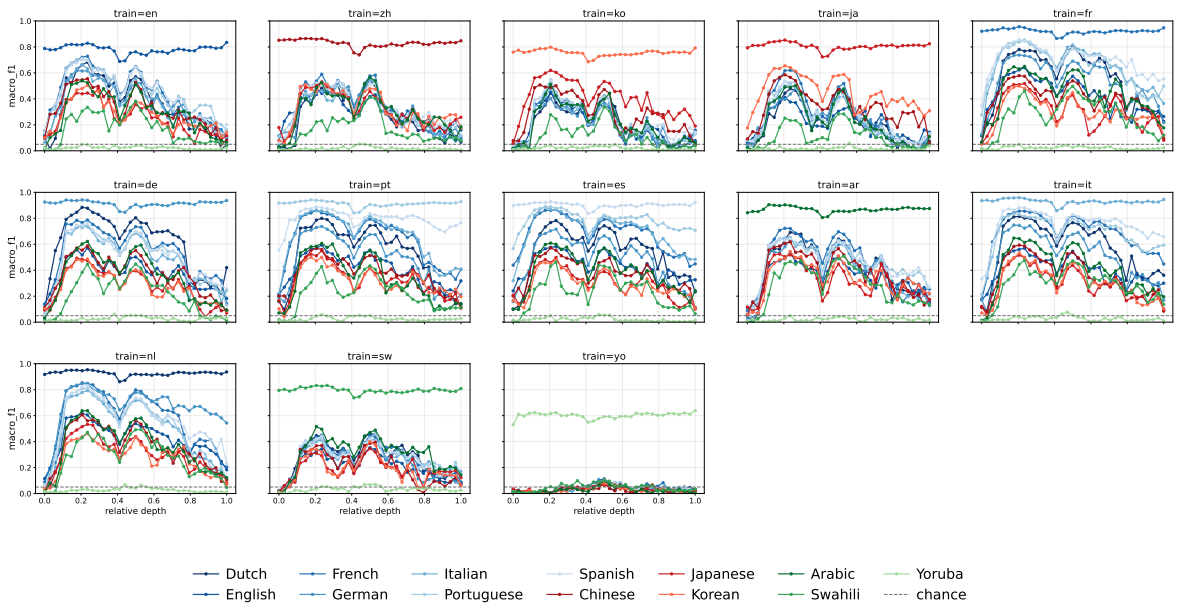


Figure 16: Layerwise cross-lingual emotion probe transfer in **Gemma 4 E2B IT**. Each panel uses one training language; each line shows macro-F1 on one test language. The dashed line indicates chance performance (≈ 0.05).

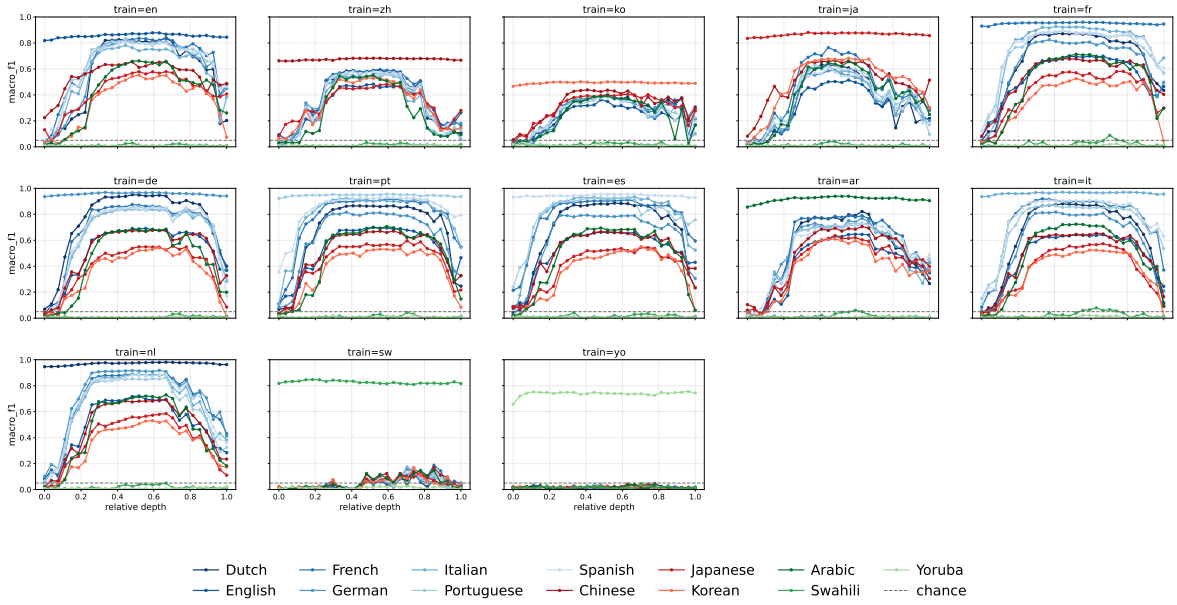


Figure 17: Layerwise cross-lingual emotion probe transfer in **Qwen 2.5 7B**. Each panel uses one training language; each line shows macro-F1 on one test language. The dashed line indicates chance performance (≈ 0.05).

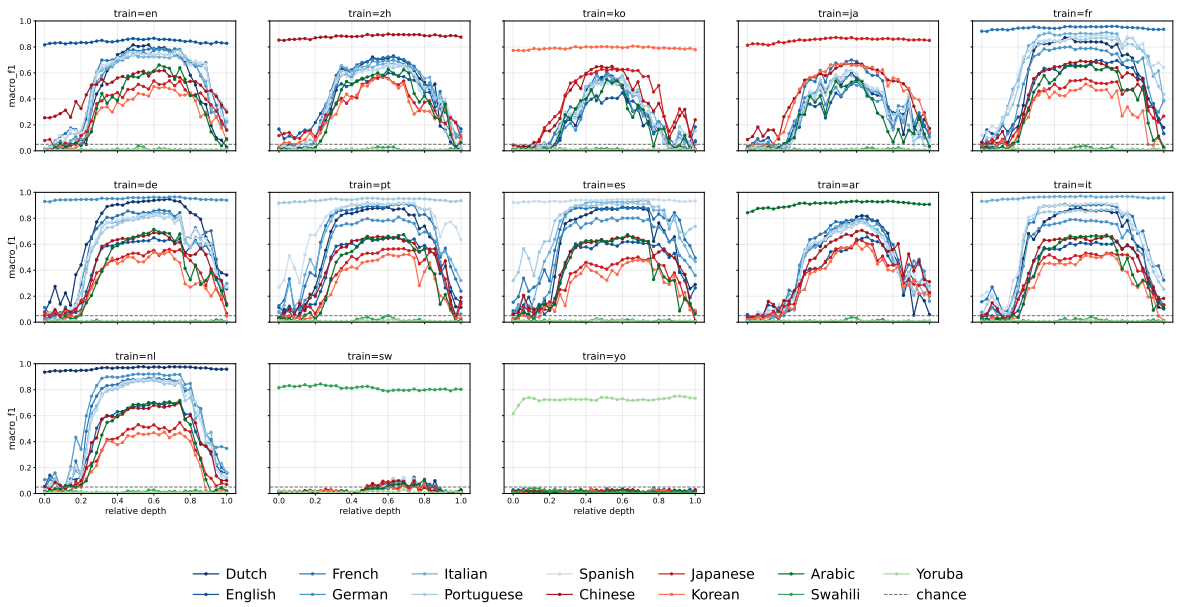


Figure 18: Layerwise cross-lingual emotion probe transfer in **Qwen 2.5 3B**. Each panel uses one training language; each line shows macro-F1 on one test language. The dashed line indicates chance performance (≈ 0.05).