

Leaving the English Basin: A Logit-Lens View of Cross-Lingual Emotion Transfer in Multilingual LLMs

Zarina Uvalieva

AIRUN

zarina.uvalievaa@gmail.com

Abstract

We study how multilingual LLMs encode emotion in English, Russian, and Kyrgyz—a low-resource Turkic language absent from prior emotion-probing work—on a parallel corpus of 1,480 sentences with six Ekman labels mapped from GoEmotions. Across four decoder LLMs (Gemma 4, Qwen3, Llama-3.1, Mistral), an encoder baseline (XLM-R), and an expanded 8-model sample, we combine layer-wise probing, similarity metrics (CKA, Procrustes, RSA), leakage-free cross-lingual transfer, and a null-calibrated sentence-level logit lens. Three findings. (1) What separates languages in the “latent-English” pivot (Wendler et al., 2024) is not the mid-network English level—which a shuffled-representation null reproduces—but the depth of the *exit* from the English basin: on pooled sentence representations Kyrgyz exits only in the last third of every applicable decoder ($d_{.5} \approx .97/.86/.97$) and never earlier than Russian ($.97/.64/.75$; under a last-token lens both languages tie late, $.92-.97$)—a lossier pivot, consistent with its larger transfer gap. (2) Similarity and transfer *dissociate*: Mistral has the highest RU–KY CKA (.800) yet the lowest KY transfer (.239), so CKA is an unreliable proxy for cross-lingual generalization. (3) Encoder/decoder architecture does not order Kyrgyz transfer: XLM-R leads (.384) but mBERT is second-to-last (.198). XLM-R’s edge is not significant on absolute KY ($p=.081$); it concentrates in the into-Kyrgyz directions (interaction $p<.01$ vs. Qwen3) and reaches significance only under a secondary chance-corrected metric—consistent with pre-training data scale and domain rather than architecture (an interpretation, not a ranking predictor). All 8 models beat chance in every language.

1 Introduction

Modern multilingual LLMs are typically evaluated on high-resource languages, leaving open two questions central to multilingual representation learning:

(a) what drives cross-lingual transfer to typologically distant low-resource languages, and (b) how the internal geometry of these representations compares across architectures and pre-training regimes. Emotion is an ideal probe for both: cognitively universal yet lexically language-specific, with signal at lexical and contextual levels—yet how multilingual LLMs encode it remains poorly understood.

In this work, we address these gaps through a systematic probing study of four decoder-only LLMs—**Gemma 4 E4B** (multimodal, Per-Layer Embeddings; Gemma Team, Google DeepMind 2026), **Qwen3-8B** (Qwen Team 2025), **Llama-3.1-8B** (Meta AI 2024), and **Mistral-7B-v0.3** (Jiang et al. 2023)—and one encoder baseline, **XLM-R-large** (Conneau et al., 2020), on a parallel emotion corpus spanning **English** (EN), **Russian** (RU), and **Kyrgyz** (KY), the latter a low-resource Turkic language with limited NLP tooling.

Our contributions are:

1. A *null-calibrated English-pivot* result: a sentence-level logit lens with a shuffled-representation null shows that the informative signature is the *exit* from the English basin, not the raw English decoding level. *Kyrgyz leaves the basin only in the last third* in all three decoders the lens applies to (Gemma 4’s PLE precludes the fourth), *and never earlier than Russian*, whose exit depth is model-specific (pooled representations; a last-token lens yields late ties, $.92-.97$). This complements Wendler et al. (2024) with a re-emergence–depth account of the transfer gap (Section 6.4).
2. A direct demonstration that *representational similarity and functional transfer dissociate*: among the four main decoders, Mistral achieves the highest RU–KY CKA (.800) yet the *lowest* KY_{avg} transfer (mean accuracy over the four KY-involving directions;

.239; Llama vs. Mistral on EN→KY +.159, $p < .001$), showing CKA is an unreliable proxy for cross-lingual generalization (Section 6.2).

3. A *leakage-free transfer protocol* for parallel corpora—a methodological caution in its own right: with parallel sentences, an uncontrolled split lets transfer exceed within-language probing (leaky XLM-R RU→EN .615 vs. .551 within; leakage-free .476)—and under it (disjoint splits, nested layer selection), evidence that *encoder/decoder architecture does not order the 8 models on Kyrgyz transfer*: XLM-R leads while mBERT is second-to-last, with five decoders between the two encoders; the edge concentrates into-KY (significant vs. Qwen3), and we read XLM-R’s outlier status through pre-training corpus scale (predictive, not causal; Section 6.3).

All experiments run on a new parallel EN/RU/KY emotion corpus built from GoEmotions (Demszky et al., 2020)—LLM-translated (Claude 3.7 Sonnet), human-audited for translation fidelity, and checked against an independent human-translated control set (Section 3.3, Appendix L)—which we release as a resource together with all code. A secondary depth trend and pooling ablation are in Appendix M; we treat the depth trend as a hypothesis.¹

2 Related Work

Probing classifiers. Alain and Bengio (2017) introduced linear probes as a tool for interpreting neural representations. Tenney et al. (2019) showed that BERT rediscovers classical NLP pipeline stages at different depths. Hewitt and Liang (2019) highlighted that probing accuracy alone can be misleading and proposed *control tasks* and *selectivity* as diagnostics; we adopt selectivity-style diagnostics (a shuffled-label check plus a lexical baseline). Pimentel et al. (2020) and Voita and Titov (2020) reframe probing information-theoretically; Belinkov (2022) synthesizes the debate. We follow this paradigm, extended to generative decoders.

Emotion representation in NLP. Emotion analysis in NLP has a long history grounded in psychological taxonomies, most notably Ekman’s six basic emotions (Ekman, 1992). Prior dataset work includes EmoBank (Buechel and Hahn, 2017), NRC

affect intensities (Mohammad, 2018), GoEmotions (Demszky et al., 2020), the multilingual XED corpus (Öhman et al., 2020), and BRIGHTER—human-annotated emotion data for 28 languages (Muhammad et al., 2025a), deployed in SemEval-2025 Task 11 (Muhammad et al., 2025b), covering Russian and Tatar but not Kyrgyz (hence our parallel corpus). We do not use its Russian portion here (multi-label intensity scheme and different domains vs. our single-label Ekman-6 design); it is a natural external validation set for follow-up. From a representational angle, Triantafyllopoulos et al. (2022) show that probing accuracy for affective content varies by layer and architecture in speech transformers. Nanda et al. (2023) show emergent linear representations of structured concepts in sequence models. Plaza-del Arco et al. (2024) survey the field and identify multilinguality and interpretability as key gaps. We extend these lines of inquiry to emotion probing in multilingual generative LLMs.

Cross-lingual similarity and transfer. Kornblith et al. (2019) introduced CKA for comparing neural network representations; subsequent work cautions that CKA is dominated by high-variance directions and can be insensitive to functionally important differences (Ding et al., 2021; Davari et al., 2023), motivating our use of transfer probing as a behavioral complement. Wu and Dredze (2019) find language-agnostic middle layers in mBERT; Conneau et al. (2020) extend this to XLM-R. More recently, Wendler et al. (2024) argue that multilingual decoder LLMs may compute in a latent “English” representation space; Dumas et al. (2024) separate language from concept via patching, and Schut et al. (2025) still find English-shaped decision spaces in open-ended settings. A complementary line localizes multilinguality in *language-specific neurons* (Tang et al., 2024) or layer-wise pipeline stages (Zhao et al., 2024); our exit measure is representation-level, orthogonal to neuron-level localization. Our pooled lens adds a low-resource Turkic target with a null calibration. Recent work argues pre-training data composition may matter more than architecture for low-resource transfer (Chirkova and Nikoulina, 2024; Üstün et al., 2024), motivating our encoder/decoder comparison across divergent pre-training mixes.

Low-resource languages. Low-resource and Turkic NLP remain under-served (Joshi et al., 2020); exceptions include large-scale Turkic MT (Mirzakhlov et al., 2021) and handling lan-

¹Code and data: <https://github.com/Zarina33/leaving-the-english-basin>

guages unseen by multilingual encoders (Muller et al., 2021). Our inclusion of Kyrgyz—a Turkic language with limited NLP resources and, to our knowledge, no prior emotion-probing study—extends this inquiry to generative LLMs and affective content.

3 Data

3.1 Source Dataset

We use **GoEmotions** (Demszky et al., 2020), mapping its 28 fine-grained labels to 6 basic Ekman emotions (anger, disgust, fear, joy, sadness, surprise) via the standard Ekman grouping. We keep only *single-label* examples (dropping multi-label and neutral) and sample 250 per class (1,500 total). After translation validation, the final corpus contains **1,480 parallel sentence triples** (1.3% excluded due to quality issues).

3.2 Parallel Translation

Russian and Kyrgyz translations were generated using Claude 3.7 Sonnet (Anthropic, 2025) with prompts emphasizing emotional content preservation over literal translation. Translation quality was assessed by trilingual (EN/RU/KY) native speakers, who judged source–target faithfulness in both directions. A primary annotator checked all 1,500 translated sentence triples and corrected or excluded examples with semantic drift, yielding the final 1,480. An independent second trilingual annotator then validated a stratified subsample of 200 sentences on a 4-point scale (OK / Minor edit / Major edit / Reject). Both annotators rated >90% of translations as OK (marginal OK-rates 96.5% and 91.0%—the inputs to AC1), with 0 rejections (the subsample postdates the primary pass and its 1.3% exclusions) and raw agreement of 88.5%. Cohen’s κ is low (0.035) from the prevalence paradox; Gwet’s AC1 (Gwet, 2008) corrects for it: 0.880 (“almost perfect”).

3.3 Human-translated Control Sample

To test whether probing reflects emotion encoding rather than similarity to Claude’s output distribution, we built a human-translated control set. A stratified sample of 102 source EN sentences (17 per class) was re-translated into Russian and Kyrgyz by an independent trilingual EN/RU/KY native translator blind to the Claude versions and barred from MT/LLM tools; after a quality-review pass, the Claude-vs-human comparison is in Appendix L.

3.4 Corpus Statistics

Table 1 summarizes the final corpus. All three languages contain 1,480 parallel sentences across six emotion classes (230–250 per class; Table 1), with comparable average token lengths (13.8–14.2 tokens), indicating that the translation process did not substantially alter sentence length or structure.

Language	Sentences	Avg. tokens	Script
English	1,480	14.2	Latin
Russian	1,480	13.8	Cyrillic
Kyrgyz	1,480	14.1	Cyrillic

Table 1: Corpus statistics. Six emotion classes: 250 examples each, except anger (230; all 20 exclusions fell in this class).

4 Models

We study four decoder-only LLMs loaded in 4-bit NF4 quantization (Dettmers et al., 2023) and one multilingual encoder baseline (XLM-R-large, fp16). Table 2 summarizes their key properties.

Model	Params	Layers	Hidden	Type
Gemma 4 E4B	4.5B*	42	2,560	PLE
Qwen3-8B	8B	36	4,096	Dense
Llama-3.1-8B	8B	32	4,096	Dense
Mistral-7B	7B	32	4,096	Dense
XLM-R-large	550M	24	1,024	Encoder

Table 2: Model overview. *Effective parameters (8B total): Per-Layer Embeddings (PLE), a secondary embedding table applied at each decoder layer, account for the difference. All models include the embedding layer (layer 0), giving 43, 37, 33, 33, and 25 total layers for analysis, respectively.

Decoder LLMs. **Gemma 4 E4B** (Gemma Team, Google DeepMind, 2026): multimodal, 4.5B effective parameters (8B total) via Per-Layer Embeddings (used text-only). **Qwen3-8B** (Qwen Team, 2025): dense 8B decoder. **Llama-3.1-8B** (Meta AI, 2024): dense 8B, strong multilingual coverage. **Mistral-7B-v0.3** (Jiang et al., 2023): dense 7B with grouped-query and sliding-window attention. Each fits in 5–6 GB VRAM under NF4.

XLM-R-large (Conneau et al., 2020) is a multilingual encoder (550M parameters, 24 layers) pre-trained with multilingual MLM on CommonCrawl in 100 languages—monolingual streams, no parallel data or cross-lingual objective; a non-generative baseline.

These five models form the basis of the main analyses; we extend the sample to $N = 8$ in Appendix P (mBERT, Qwen2.5-7B, OLMo-2-7B) for robustness. Hidden states are extracted as mean-pooled representations across non-padding tokens for all models. We use mean rather than last-token pooling because every decoder is run on the raw sentence, with no prompt or chat template: there is no summary token analogous to `[CLS]`, and the final token of an unprompted sequence is just the last lexical item. (Llama-3.1, Mistral, Qwen2.5 and OLMo-2 are base checkpoints; Gemma 4 E4B and Qwen3-8B are the released post-trained checkpoints, `gemma-4-E4B-it` and `Qwen3-8B`.) An ablation under last-token pooling (Appendix M) shows uniformly lower probing accuracy across all 12 (model, language) cells (mean drop 0.13, range 0.081–0.220; XLM-R excluded, as last-token pooling is not meaningful for a bidirectional encoder), so we adopt mean pooling throughout.

5 Methodology

Layer-wise probing. For each layer and language, we train a linear classifier (single-layer PyTorch, CrossEntropyLoss, Adam, $\text{lr}=0.01$, weight decay 10^{-4} , 300 epochs) with 5-fold stratified cross-validation, reporting mean accuracy $\pm$ std; best-layer significance via permutation (10,000 shuffled-label runs; $p = (1 + \#\{\text{perm} \geq \text{real}\}) / (1 + 10,000)$).

Representational similarity. We compute linear CKA (Kornblith et al., 2019) per layer between language pairs (≈ 1 = similar geometry), plus Procrustes and RSA (definitions in Appendix C); cluster geometry uses cosine silhouettes and UMAP (Appendix A).

Cross-lingual transfer (leakage-free). Because the corpus is parallel, training on all of language A and testing on all of language B evaluates the probe on *translations of its training sentences*—a leak. We instead split sentence indices once into disjoint train/dev/test sets (60/20/20; identical across languages so parallel triples stay together), train on the source-language train split, select the best layer on the *target-language dev split* (sentences disjoint from both train and test; target labels enter only this layer choice, never training), and report accuracy on the target-language test split (nested selection). This removes both the parallel-sentence leak and the best-layer-on-test bias; left uncontrolled,

the leak can push transfer accuracy above within-language probing (Section 1). Pairwise significance uses the same paired bootstrap and Holm correction as for within-language probing. Throughout, “functional transfer” is meant in this probe-level sense: a readout trained on one language works on another’s representations—a behavioral test of the *representations*, as opposed to geometric similarity, and not evidence that the model itself uses this information when generating text (see Limitations).

Per-emotion analysis. Binary one-vs-rest probing for each of the 6 emotions with class-weighted cross-entropy to handle the $\approx 1:5$ imbalance. We report *macro-F1* (rather than accuracy) at the best layer to avoid inflation from majority-class bias.

Probe validity. Following Hewitt and Liang (2019), we run two controls to ensure probes capture genuine emotion structure rather than memorizing vocabulary or exploiting representational artifacts: (a) a *shuffled-label* check—control accuracy must fall to the 1/6 floor; *selectivity* equals real accuracy minus control accuracy (weaker than Hewitt–Liang’s learnable control; the TF-IDF baseline covers vocabulary memorization); and (b) a 2-layer *MLP probe* (hidden=256, ReLU, dropout=0.1) to test whether linear probes underestimate encoded information.

Statistical testing. For each model-language pair, we compute 95% bootstrap confidence intervals (10,000 resamples over the 1,480 per-example predictions from 5-fold CV). Pairwise model differences use a bootstrap test on the accuracy difference (10,000 resamples) with Holm–Bonferroni correction within each language, and likewise within each transfer direction for the decoder pairwise transfer comparisons (Appendix K). A permutation test (10,000 shuffled-label runs) at each model’s best layer from the initial probing run (see Limitations) confirms above-chance performance (all 15 cells $p=1 \times 10^{-4}$, i.e. 0 of 10,000).

6 Results

6.1 Layer-wise Probing Accuracy

Table 3 shows best-layer probing accuracy (per-layer curves in Figure 3, Appendix M); all 15 main-model cells are significant under permutation ($p=1 \times 10^{-4}$; chance 1/6).

All models decode emotion well above chance in every language; on Kyrgyz the decoders cluster

Model	EN			RU			KY		
	L	Acc	CI	L	Acc	CI	L	Acc	CI
Gemma 4	0	.603	.58–.63	7	.508	.48–.53	4	.453	.43–.48
Qwen3	10	.605	.58–.63	11	.582	.56–.61	9	.482	.46–.51
Llama	1	.625	.60–.65	19	.568	.54–.59	3	.478	.45–.50
Mistral	5	.611	.59–.64	9	.570	.54–.59	5	.451	.42–.48
XLM-R	11	.551	.53–.58	21	.482	.46–.51	14	.459	.43–.48

Table 3: Best layer-wise probing accuracy with 95% bootstrap CIs (10k resamples over 1,480 examples), from a single unified probing run. L = best layer. All model×language cells are well above chance (1/6); on English Llama has the highest point estimate (.625) and the encoder XLM-R the lowest (.551), but *no decoder pair is statistically distinguishable* (none survives Holm correction; Appendix G; no formal test is run against XLM-R here); on Kyrgyz the four decoders cluster tightly (.451–.482, overlapping CIs).

tightly (.451–.482, overlapping CIs), so fine KY differences are not robust. The four main decoders drop 12–16 pp from EN to KY, the two multilingual encoders *least* (8–9 pp; full 8-model range 8–17 pp; Table 21). The small drops are not predictive: mBERT’s mostly reflects its low EN start (.482), and mBERT ends second-to-last on transfer (Section 6.3).

We also observe a *depth trend*—EN peaks at Gemma 4/Llama L0/L1, Mistral L5, Qwen3 L10, XLM-R L11: a continuous spread in L/N (.00–.46), not clean early-vs-late clusters. The pattern is not clean across languages (Llama RU peaks at L19) and is pooling-dependent (last-token shifts Gemma 4 EN 0→42); a hypothesis (Appendix M).

6.2 CKA Cross-lingual Similarity

We assess cross-lingual similarity using three complementary metrics: CKA (Table 4), orthogonal Procrustes, and RSA. For EN–RU, Llama achieves the highest similarity on CKA and Procrustes (CKA .830, Procrustes .915; Tables 4 and 8); on RSA, Gemma 4 edges Llama (.884 vs .878); Mistral is second on CKA/Procrustes (.794/.890), third on RSA (.851). The metrics also diverge on *depth* for Llama (CKA L31, a plateau $\geq .82$ from L1; Procrustes L2, RSA L14); the three capture different properties (kernel similarity, rotational alignment, rank-order preservation), and none is definitive alone.

Qwen3’s metrics disagree too (CKA/Procrustes L19–22 vs. RSA L6–8, spread $\approx .44$); the caveat applies to all four.

For RU–KY (both Cyrillic), Procrustes shows high early-layer similarity for Mistral (plateau .879 from L2, peak L7) and Llama (L2), consistent with Cyrillic overlap at the embedding level; Llama’s EN–KY also peaks early (L2), Mistral’s

deeper (L7). Gemma 4 peaks at L0 on all pairs on Procrustes, matching its early probing profile. Peak CKA for the remaining four models is in Table 23 (Appendix Q): XLM-R’s peaks (.68–.74) are unremarkable—below Llama’s and Mistral’s on every pair—yet XLM-R leads KY transfer, reinforcing the dissociation established below.

6.3 Cross-lingual Transfer

Transfer is measured under the leakage-free protocol of the Methodology section (disjoint train/dev/test splits, nested target-dev layer selection); uncontrolled, the leak can push transfer above within-language probing.

On absolute KY transfer, XLM-R leads (KY_{avg} .384), ahead of the best decoder (Qwen3 .349; $\Delta = .036$, $p = .081$, *n.s.*), mBERT (.198) and OLMo-2 (.183) far below. After Holm, XLM-R is significantly above every model *except* Qwen3 and Llama (vs. mBERT $+ .186$, $p < .001$; all tests in Appendix H). The absolute-KY margin is thus small and statistically indistinguishable at $n = 296$; a matched-precision (fp16) Qwen3 rerun leaves it so (gap $+ .015$, *n.s.*; Appendix N). The picture is not a clean encoder–decoder split: mBERT, the second encoder, is second-to-last. The strongest, *ratio-free* evidence is directional: XLM-R’s advantage is concentrated in the *into*-Kyrgyz directions (EN→KY/RU→KY: .392/.358 vs. the per-direction best decoder, Llama, .358/.304), while on KY-source it is matched or exceeded (.409/.378 vs. Qwen3 .443/.382; the envelope switches models—see the fixed-model test below). What XLM-R does better is *represent Kyrgyz as a probe target* (lowest KY fragmentation, 2.03; Table 24); no normalization involved. The asymmetry is significant against Qwen3, the decoder that beats XLM-R on KY-source (DiD $+ .108$, $p = .008$; fp16 $+ .074$,

Pair	Gemma 4		Qwen3		Llama		Mistral	
	CKA	L	CKA	L	CKA	L	CKA	L
EN→RU	.599	12	.604	22	.830	31	.794	31
EN→KY	.575	12	.620	19	.713	12	.736	18
RU→KY	.599	21	.620	19	.731	2	.800	1

Table 4: Peak linear CKA and layer (L) at which it peaks, for the four main decoders; the remaining four models are in Table 23. Llama achieves the highest EN→RU alignment (.830); Mistral achieves the highest RU→KY (.800).

Direction	Gemma 4	Qwen3	Llama	Mistral	XLm-R
EN → RU	.497	.581	.537	.537	.470
EN → KY	.250	.294	.358	.199	.392
RU → EN	.513	.544	.513	.527	.476
RU → KY	.290	.277	.304	.240	.358
KY → EN	.321	.443	.341	.280	.409
KY → RU	.307	.382	.361	.236	.378
KY _{avg}	.292	.349	.341	.239	.384
η_{cc}	.371	.460	.487	.198	.710

Table 5: Leakage-free cross-lingual transfer accuracy (source-trained probes; disjoint sentence splits, target-dev layer selection, held-out target test—“zero-shot” for probe training, with target labels used only for the layer choice). XLm-R leads absolute KY (KY_{avg} .384; vs. Qwen3 .349, $p=.081$ n.s.; Appendix H). The last row, chance-corrected efficiency $\eta_{cc} = (KY_{avg} - 1/6) / (EN \leftrightarrow RU - 1/6)$, where $EN \leftrightarrow RU$ denotes the mean of the EN→RU and RU→EN accuracies, controls for model strength above the 1/6 chance floor. Under η_{cc} , XLm-R significantly leads every decoder including the best ones (.710 vs. Llama .487, $p<.01$; vs. Qwen3 .460, $p<.001$; Appendix J). This table shows the 5 main models; the added mBERT / Qwen2.5 / OLMo-2 are in Table 22; see Section 6.3 for why mBERT’s wide η_{cc} CI leaves its ranking unresolved.

marginal); against Llama it is direction-uniform (Appendix H). A chance-corrected efficiency $\eta_{cc} = (KY_{avg} - 1/6) / (EN \leftrightarrow RU - 1/6)$ ($EN \leftrightarrow RU$ = mean of the two directions) corroborates: XLm-R *significantly* leads the top decoders (.710 vs. .487/.460, $p<.01/p<.001$; Appendix J), which absolute KY could not distinguish. We treat η_{cc} as *secondary*: any such ratio divides by an English-anchored quantity on which XLm-R is weakest (analysis in Appendix J; near-chance mBERT is additionally unstable: $EN \leftrightarrow RU$.082 above floor, CI [.11, .85]). The robust reading: XLm-R is a significant outlier; encoder/decoder does not order the 8 models. What distinguishes XLm-R is pre-training *corpus scale* (~ 2.5 TB of filtered CommonCrawl vs. mBERT’s ~ 21 GB of Wikipedia; both encoders are MLM-trained over $\gtrsim 100$ languages, so the contrast is about data scale and domain, not objective). An XLm-R-base control decomposes the remaining confound into two steps. mBERT→base is the data step, though not a pure swap (178M→270M parameters, 119k→250k vocabulary, no NSP; Wikipedia→CommonCrawl): it adds +.105 KY_{avg} (paired bootstrap [+ .07, +.14], $p<.001$). base→large is the clean step—same corpus, same 250k vocabulary, same objective, only

scale changes ($2 \times$ depth, $1.3 \times$ width)—and adds +.081. On absolute KY, where both steps are precisely estimated, the two contribute comparably, and XLm-R-large’s lead is their joint effect: base alone (.303) sits mid-pack, between Gemma 4 (.292) and Llama (.341). Under η_{cc} the data step accounts for nearly the whole mBERT→large gap (+.277 of +.320), but that difference is not resolvable at $n=296$ (CI [−.20, +.60], mBERT’s near-chance denominator; Appendix J), so we do not read data dominance off the ratio (Table 22). We frame this as an interpretation of XLm-R’s outlier status, not a ranking predictor (predictive, not causal; Section 7).

6.4 The Exit from the English Basin

Does high $EN \leftrightarrow RU$ transfer reflect language-agnostic representations or an English pivot (Wendler et al., 2024)? We test this directly with a sentence-level *logit lens*: for each layer we project the mean-pooled hidden state through the model’s final norm and unembedding, and read the language of the top-20 decoded tokens (applied to the three dense decoders; Gemma 4’s Per-Layer Embeddings preclude the lens; Appendix O). Figure 4 plots the raw top-20 fractions (English / input-

language Cyrillic; a third class, “other”, absorbs the rest); in-text numbers renormalize over the word-like classes, EN/(EN+Cyrillic). For both non-English inputs the mid-network layers decode predominantly to English—a level the English-biased unembedding reproduces even for shuffled vectors (Appendix O)—so the informative signature is the *exit*, which differs by language and model. We operationalize the exit as $d_{.5}$ (depth where the English share drops below .5 for good) and $AUC_{[.5,1]}$ (its mean over the second half). Kyrgyz *leaves the basin late in every model*— $d_{.5} = .97/.86/.97$ (Llama/Qwen3/Mistral), all in the last third, in no model earlier than Russian. Russian’s exit depth is model-specific: Llama holds English to the final layer (.97) then switches abruptly; Qwen3 switches at two-thirds (.64); Mistral descends from $\sim .6$ (.75). The robust contrast is thus the *depth ordering*—Kyrgyz never exits first, and always in the last third—rather than the manner of exit. The ordering is strict in Qwen3/Mistral, a tie in Llama (both .97)—a consistent direction, not an effect size—and survives upper-bound fragment reattribution ($d_{.5}$ shifts $\leq .03$, $KY \geq RU$ intact), a last-token lens (ties .92–.97, same AUC signs; Appendix O), and sentence-level bootstrap (2,000 resamples: every $d_{.5}$ unchanged at the layer grid (1/32 or 1/36); $KY \geq RU$ in 100% of resamples per model). Manner varies by model: $AUC_{[.5,1]}$ (KY/RU) favours Kyrgyz in Qwen3 (.62/.24) and Mistral (.77/.51) but Russian in Llama (.74/.90)—which language holds English longer is not fixed across models (each gap tight under sentence-level bootstrap; a text-specified last-third window reproduces the ordering, and within Mistral RU falls .77 vs. .22 for KY over .56–.84—amplitude, not abruptness; Appendix O). We read Kyrgyz’s consistently late departure as a lossier pivot—the network holds it in the English basin until the final layers, leaving less depth for language-specific processing—consistent with its larger transfer gap (Table 5). The reading is cross-language, not a per-model dose–response: RU exit depth varies across models with no matching RU-transfer variation (Qwen3’s early switch costs no RU transfer), so depth is not a graded within-language predictor. Rather than extending Wendler et al.’s mid-network evidence—which a pooled lens cannot certify—we measure the *re-emergence depth* of the input language and find Kyrgyz’s consistently late. “English basin” is accordingly an operational label for the depth range over which the lens decodes to English tokens: the

lens reads the language of decoded tokens, and does not show that the model computes in English or that these states are English semantic states. The pooled vector is, deliberately, the same aggregate our probing and transfer analyses operate on, so the pooled exit is the quantity relevant to the transfer-gap reading; the last-token ties qualify its per-token interpretation, not that link (Appendix O).

6.5 Probe Validity

Shuffled-label control accuracy is near chance (12 decoder cells .149–.191; XLM-R .17–.18), selectivity .29–.46 (XLM-R .29–.37; Appendix D)—probes decode genuine structure, not artifacts. MLP probes provide only marginal gains (mean gap +.016), indicating predominantly linear encoding. The shuffled-label control rules out a random mapping but not emotion-bearing *vocabulary*; a TF-IDF baseline (1–2 grams + logistic regression, same 5-fold CV) reaches .545/.451/.470 on EN/RU/KY. The best-layer probe exceeds it by +.080 (EN) and +.131 (RU); on Kyrgyz only Qwen3 (.482) and Llama (.478) marginally exceed it, while Gemma 4, Mistral, and XLM-R fall *below* (.451–.459 vs. .470); at gaps of +.012/+.008 obtained by selecting the best of 33–37 layers (without nested selection; Limitations), exceeding the bar is not established for any model. This concerns emotion decodability only—the lens classifies token *language*, not emotion content—so the exit-depth ordering (Section 6.4) does not rest on beating the lexical baseline. We therefore read the KY probing numbers as *decodability* claims, not claims about the depth or quality of Kyrgyz *representations*—consistent with the early peaks (Section 6.1) and the weaker Kyrgyz transfer (Section 6.3). Decodability and transferability are distinct: XLM-R’s KY probing sits *below* the TF-IDF bar (.459 vs. .470) yet it leads KY transfer—not richer decodable signal, better cross-language alignment. Under the same leakage-free protocol, TF-IDF transfers at $\sim .13$ from an EN source and $\sim .20$ –.25 from RU/KY sources (Latin residue, RU–KY borrowings; Appendix F); the *best model per direction* clears the bar on every KY direction (+.14–.26), OLMo-2 sits at or below it on three of four, mBERT on two, Mistral within .03 outside EN→KY.

Accuracy decreases with length for all four decoders (Gemma 4: .649 at 1–5 words vs. .531 at 21+; Table 10), consistent with mean-pooling dilution (monotone for Gemma 4/Llama, small last-bin uptick for Qwen3/Mistral); the shared pattern does

Emotion	Gemma 4 E4B			Qwen3-8B			Llama-3.1-8B			Mistral-7B		
	EN	RU	KY	EN	RU	KY	EN	RU	KY	EN	RU	KY
Anger	.584	.613	.605	.583	.602	.590	.608	.587	.637	.616	.593	.625
Disgust	.655	.643	.645	.674	.661	.695	.668	.684	.695	.711	.666	.685
Fear	.764	.711	.688	.777	.712	.678	.775	.716	.694	.774	.720	.685
Joy	.814	.801	.748	.822	.825	.754	.794	.772	.778	.796	.807	.778
Sadness	.696	.648	.660	.682	.678	.669	.680	.691	.688	.698	.675	.684
Surprise	.624	.639	.623	.643	.605	.626	.662	.607	.643	.651	.642	.633

Table 6: Best one-vs-rest *macro-F1* per emotion, language, and model; macro-F1 accounts for the $\approx 1:5$ class imbalance in OvR probing. Bold marks the best emotion per column. **Joy** remains the most linearly separable emotion; **Fear** ranks second on EN and RU, while on KY **Disgust** ties or overtakes it in three of four decoders (see text for the persisting KY>EN cells). XLM-R is omitted: the class-weighted OvR pipeline was not re-run on it (Appendix B). Numbers here (and in Table 7) come from the separate class-weighted OvR pipeline on the same extracted hidden states (script provenance in the repository README); the unified probing re-run does not alter them.

not differentially explain early peaking.

6.6 Per-emotion Analysis

Joy achieves the highest macro-F1 (>0.75) across nearly all model-language combinations (Table 6), consistent with its lexical distinctiveness. **Fear** ranks second on EN and RU (.71–.78); on KY, **Disgust** ties or overtakes it in three of the four decoders (Table 6). **Anger** is hardest (.58–.64). Under macro-F1, Joy/Fear/Surprise decrease from EN to KY across models, but the KY>EN pattern persists for **Anger** in all four decoders, for **Disgust** in two (Qwen3, Llama), and for **Sadness** in one (Llama)—possibly strong lexical marking in Kyrgyz. Multiclass confusions (Appendix S) show *which* categories blur: Anger is the hub (its pairs with Disgust, Surprise and Sadness are the three most confused on the cross-language mean, .14–.16 of the pair’s sentences), Fear–Joy the cleanest (.04), and the ranking of the 15 pairs is shared across languages (Spearman .78–.97 in every model). Confusion is not uniform within negative valence (Anger–Fear .07 vs. Anger–Disgust .16), so the probes separate more than a positive/negative split. UMAP projections at the best probing layers (Appendix A, Figure 2) show no geometric clustering despite above-chance decodability, consistent with the near-zero silhouettes.

7 Discussion

Depth trend and lexical signal. Best-probing layers form a continuous depth spread, seed- and pooling-dependent (Section 6.1, Appendix M)—a hypothesis, not a finding. The TF-IDF baseline quantifies the early-peak reading; OLMo-2’s independent EN L0 peak shows it is not Gemma-4-

specific.

Cross-lingual transfer and English pivot.

Leakage-free EN $\leftrightarrow$ RU transfer is highest for Qwen3 (mean .56; Mistral and Llama both .53); on Kyrgyz no decoder pair survives Holm on probing (Appendix G), and only KY-involving directions yield significant decoder–decoder transfer effects (Appendix K). The logit lens (Section 6.4) recasts Wendler et al.’s “latent English” account in re-emergence–depth terms: a pivot Kyrgyz leaves only in the last third and never before Russian, whose exit depth is model-specific.

Encoder/decoder does not predict Kyrgyz transfer.

A natural reading attributes XLM-R’s KY lead to encoder architecture; three lines argue otherwise (Section 6.3). *Absolute KY*: “encoder” alone confers nothing—mBERT sits significantly below five of six decoders (Appendix I). The *directional asymmetry* against Qwen3 (marginal at fp16) suggests, normalization-free, that the edge is about representing Kyrgyz; against Llama the advantage is direction-uniform. *Chance-corrected efficiency* corroborates (XLM-R above every decoder; denominator caveat in Appendix J; mBERT unresolved). What distinguishes XLM-R is pre-training data *scale and domain* (~ 2.5 TB CommonCrawl vs. ~ 21 GB Wikipedia); an XLM-R-base control supports the decomposition (matched-depth data swap +.105, same-data depth doubling +.081; Section 6.3); at $n=8$ an interpretation (predictive, not causal), not a ranking predictor—nor does disclosed *breadth* suffice (Gemma 4: 140+ languages (Gemma Team, Google DeepMind, 2026), KY membership undisclosed, yet mid-pack at .292). Tokenizer fertility is statistically indistinguishable

from zero at $n=8$ ($r=-.46$, $p=.25$; Table 24) and unreliable alone (mBERT: low fragmentation, near-bottom transfer).

Mistral’s Kyrgyz deficit: CKA–transfer dissociation. Among the four main decoders, Mistral has the highest RU–KY CKA (.800, L1) yet the lowest average KY transfer (KY_{avg} .239)—significant vs. both top decoders on EN→KY and vs. Qwen3/Llama on several KY-source directions (Table 16). High embedding similarity (likely shared Cyrillic) does not guarantee transfer; probing it directly is the necessary complement.

8 Conclusion

First (3 decoders): a null-calibrated lens locates the signal in the *exit* from the English basin—on pooled representations Kyrgyz exits only in the last third, never before Russian—consistent with its transfer gap. **Second** (4 main decoders): similarity and transfer *dissociate* (Mistral: top RU–KY CKA .800, lowest KY_{avg} .239). **Third** (8 models): encoder/decoder *does not order KY transfer*—XLM-R leads, mBERT second-to-last; the edge concentrates into-KY vs. Qwen3 ($p<.01$, fp16-marginal); the absolute gap is n.s., significant only under secondary η_{cc} .

Ethics Statement

GoEmotions is derived from Reddit comments, which skew toward English-speaking, young, male demographics. Emotion annotations may reflect cultural norms of this population; our Russian and Kyrgyz translations inherit this Western-centric framing of emotion categories. LLM-generated translations may amplify cultural and affective biases present in the models’ pre-training data. Researchers should exercise caution before applying these findings to downstream applications involving vulnerable populations, mental health, or content moderation in Kyrgyz-language contexts.

Limitations

Typology and resource level are confounded. Our two non-English languages vary along two entangled dimensions: Russian is both typologically close to English (Indo-European) and high-resource, whereas Kyrgyz is both typologically distant (Turkic, agglutinative) and low-resource. The English-pivot depth pattern is therefore consistent with a typology account, a resource account,

or both; adding a Turkic high-resource language (Turkish) would decouple them and is the single most valuable follow-up.

Our corpus is moderate—1,480 parallel sentence triples, 230–250 sentences per emotion per language—which limits statistical power for fine per-emotion claims that larger corpora would resolve. Translations were generated by Claude 3.7 Sonnet, raising the concern that probing accuracy might partly reflect similarity to Claude’s output distribution rather than emotion encoding per se. We tested this empirically on a 102-sentence human-translated control set (Section 3.3, Appendix L): across all 5 main models and both non-source languages (10 cells), the difference between probing accuracy on Claude-translated and human-translated versions of the same sentences is bounded by $|\Delta| \leq 0.059$ (mean $|\Delta| = 0.018$), with every 95% bootstrap CI including zero (at the boundary for Llama RU: [.000, +.029]; see Table 17). The LLM-translation confound, if present, is below the resolution of our experiment—and that resolution is coarse: at $n=102$ the KY-cell CIs span roughly $\pm .08$, so the control is underpowered for effects smaller than ~ 8 pp. The same check applied to transfer (EN→RU/KY, ten cells) is likewise null ($|\Delta| \leq .029$, all n.s.; Appendix L).

The translation reviewer also participated in corpus preparation, introducing potential confirmation bias; the independent second annotator (Gwet’s AC1=0.880) partially mitigates this. GoEmotions (2020) is a public dataset likely seen during pre-training of all eight models; data contamination may inflate absolute accuracy levels. Whether Kyrgyz occurs in the decoders’ pre-training corpora is undisclosed for Gemma 4, Qwen3/2.5, Llama, and Mistral (OLMo-2 documents predominantly English data); both encoders are known to include Kyrgyz. The corpus-scale interpretation therefore rests on the XLM-R–mBERT contrast, where language presence is held constant, and may conflate scale with language presence for the closed-data decoders; within that contrast itself, scale and corpus domain (CommonCrawl vs. Wikipedia) co-vary and cannot be separated at $n=2$ encoders.

Our depth profiles are pooling-dependent. We use mean pooling because decoder LLMs run on unprompted text lack a summary-token training signal, and last-token pooling yields uniformly lower accuracy in our ablation (Appendix M). However, under last-token pooling best probing layers shift substantially deeper for several models

(e.g., Gemma 4 EN: 0→42; Llama EN: 1→28). The *early-peaking* pattern observed with mean pooling therefore partly reflects emotion-bearing lexical signal aggregated from tokens dispersed across the sequence, whereas last-token probing answers a stricter “summary-at-final-token” question that no model here is explicitly optimized for on unprompted text. The main decoders also mix post-trained (Gemma 4, Qwen3) and base (Llama, Mistral) checkpoints, so model-level contrasts do not separate post-training from model family. Decoder LLMs use 4-bit NF4 while both encoders use fp16. Our Gemma 4 NF4-vs-8-bit ablation ($\Delta_{\text{acc}} \leq 0.016$; Appendix N) suggests small absolute quantization effects, and we additionally re-ran Qwen3—the decoder closest to XLM-R on absolute KY—end-to-end in fp16 (matched precision with the encoders) under the same leakage-free protocol: KY_{avg} moves from .349 to .369 ($\Delta = +.020$, 95% CI $[-.007, +.047]$, n.s.), and the XLM-R–Qwen3 gap shrinks to +.015 ($[-.025, +.054]$, $p = .46$)—the absolute-KY conclusion (XLM-R ahead, not significantly) is unchanged at matched precision (Appendix N). Gemma 4, Llama, Mistral, Qwen2.5, and OLMo-2 remain NF4-extracted (both encoders are fp16 throughout).

Per-emotion peak layers (Table 7) show high variance across languages (e.g., Gemma 4 Anger: EN=15, RU=34, KY=0), likely reflecting CV-seed sensitivity on the flat per-layer curves rather than stable architectural signatures.

Permutation tests use 10,000 runs at each model’s best layer from the initial probing run; for speed, each null run trains a lighter probe (a single 80/20 split, 60 epochs) than the real 5-fold statistic, so the test is asymmetric—with 0 of 10,000 null runs approaching the real accuracies (margins $>.2$), this cannot affect the conclusion, but the null is not an exact permutation of the reported statistic. As a check, we reran one cell—Mistral KY, the weakest real accuracy (.451) among the 15 tested—with 1,000 permutations under the exact matched protocol (5-fold, 300 epochs): null mean .166, null max .199, $p < .001$ (matched_permutation.csv). The unified re-run also moves some best layers substantially (Qwen3 KY L27→L9, XLM-R RU L12→L21) at nearly unchanged accuracy. Best-layer CIs in Table 3 do not account for selecting the maximum over 25–43 layers; the flat curves make this optimism small, and all pairwise tests fix the layer

first. Bootstrap CIs are computed over per-example predictions from 5-fold CV; within-fold correlation may lead to slight variance underestimation. The MLP–linear gap (+0.016) is reported descriptively without a formal significance test. Holm correction is applied within each test family (Appendices G–J); no correction is applied across the five families.

Probing accuracy reflects linear decodability, not functional use of emotions during generation, and our transfer results inherit this: they show that an external readout carries across languages, not that the model relies on a shared emotion representation. Establishing use requires intervening on the representation—ablating or steering along the probe directions and measuring the effect on an emotion-dependent behavior, in each language—which we leave to future work. Comparison with fine-tuned upper bounds (e.g., supervised XLM-R on GoEmotions) would help contextualize absolute accuracy levels.

Our labels are deliberately clean: single-label Ekman categories, with multi-label and neutral GoEmotions examples excluded (Section 3). The results therefore do not say whether co-occurring emotions are simultaneously decodable, nor whether the representations resolve irony and sarcasm, where surface lexical valence conflicts with the intended emotion. The second is a pointed test for our setting, since lexical signal carries much of the probing accuracy (the TF-IDF bar, Section 6.5; the mean-pooling advantage, Appendix M). The excluded multi-label examples are a ready robustness set for the existing one-vs-rest probes; literal-vs-ironic minimal pairs would need to be authored by native speakers in each language rather than translated, as irony does not reliably survive translation.

LLMs Usage Statement

During the preparation of this work, two LLMs were used in distinct roles: (1) Claude 3.7 Sonnet (Anthropic, February 2025) generated the Russian and Kyrgyz translations of the corpus (Section 3.2); (2) Claude Opus 4.8 (Anthropic, 2026) assisted with language refinement and organization of some manuscript sections. All technical content, experimental design, data analysis, and interpretation were developed by the author. Neither LLM was used for ideation of methods or generation of results.

A Cluster Geometry: Silhouette and UMAP

Figure 1 reports silhouette scores per layer for all *five* models. Across every model, language, and layer, the score is essentially zero (peak ≈ 0.004 ; most values negative), indicating that emotion classes do not form geometrically separated clusters in the cosine sense—despite linear probes achieving well above-chance accuracy (Tables 3, 5). This dissociation generalizes uniformly across encoder and decoder architectures: linear decodability (probing) and geometric cluster structure (silhouette) are distinct properties of a representation, consistent with the diffuse UMAP projections in Figure 2.

B Per-emotion Peak Layers

Table 7 reports the layer at which each emotion reaches its peak one-vs-rest macro-F1 for the four decoder LLMs (chosen by $\arg \max$ over CV-mean macro-F1, from the same class-weighted OvR pipeline as Table 6). XLM-R is omitted from the per-emotion macro-F1 comparison because we did not re-run the class-weighted OvR pipeline (Section 5, “Per-emotion analysis”) on XLM-R hidden states; its contribution to this paper is the cross-lingual transfer comparison (Table 5). Extending per-emotion analysis to XLM-R is straightforward but left for future work. **We caution that per-layer F1 curves are typically flat near the maximum (CV-mean $\arg \max$ is unstable to CV-seed on flat regions), so differences of ± 5 –10 layers should not be interpreted as substantive.** The table is provided for completeness rather than as an architectural fingerprint, and the “Kyrgyz peaks deeper” tendency described below is the only pattern we view as descriptively robust.

With that caveat, Kyrgyz emotions tend to peak at or later than their EN/RU counterparts for most emotions across models (17 of 24 cells strictly later than both), weakly suggesting that low-resource language emotion encoding may require deeper processing, though exceptions exist (Fear peaks very early on KY for Llama, L2, and Mistral, L1; Gemma 4’s KY Anger peaks at L0). Disgust peaks at nearby EN and RU layers for Gemma 4 (L10/L9) and Mistral (L3/L5), though not for Qwen3 (L1/L28) or Llama (L4/L16).

	Gemma4			Qwen3			Llama			Mistral		
	EN	RU	KY	EN	RU	KY	EN	RU	KY	EN	RU	KY
Anger	15	34	0	8	8	29	8	8	24	3	2	32
Disgust	10	9	30	1	28	36	4	16	30	3	5	26
Fear	0	15	4	5	13	34	1	29	2	2	6	1
Joy	1	6	6	27	30	23	19	4	27	7	21	24
Sadness	0	14	22	20	14	35	24	8	32	6	2	7
Surprise	15	11	23	27	17	17	2	10	27	22	4	27

Table 7: Layer at which each emotion reaches peak one-vs-rest macro-F1 (absolute layer index) for the four decoder LLMs. Kyrgyz emotions tend to peak at or later than EN/RU for most emotions, with some exceptions.

C Procrustes and RSA Peak Values

Definitions. Procrustes similarity is $1 - \|X - YR^*\|_F$ after mean-centering both hidden-state matrices, truncating to the smaller dimension, and Frobenius-normalizing, with R^* the optimal orthogonal map aligning Y to X . RSA is the Spearman correlation between the two languages’ condensed pairwise cosine-distance vectors. Both are computed per layer on a fixed 500-sentence subsample (Procrustes on 1,480 sentences is slow). Table 8 reports the peak orthogonal-Procrustes similarity and Spearman- ρ RSA over pairwise distance matrices, computed per (model, language-pair) and selected as the maximum over layers. Together with CKA (Table 4), these numerically support the Discussion claim that Llama leads EN–RU alignment on CKA and Procrustes (with Mistral second), while on RSA Gemma 4 edges Llama (.884 vs .878) and Mistral is third (.851); the three metrics also diverge on the layer at which the peak occurs—highlighting that no single similarity metric is definitive.

Metric	Pair	Gemma 4	Qwen3	Llama	Mistral
Procrustes	EN–RU	.836 (0)	.860 (19)	.915 (2)	.890 (17)
Procrustes	EN–KY	.778 (0)	.862 (19)	.864 (2)	.860 (7)
Procrustes	RU–KY	.782 (0)	.865 (19)	.841 (2)	.879 (7)
RSA ρ	EN–RU	.884 (4)	.798 (6)	.878 (14)	.851 (14)
RSA ρ	EN–KY	.835 (4)	.718 (8)	.730 (13)	.787 (2)
RSA ρ	RU–KY	.842 (4)	.768 (7)	.750 (10)	.842 (2)

Table 8: Peak Procrustes similarity and RSA Spearman ρ per model and language pair (maximum over layers; the layer of the peak in parentheses). Decoder LLMs only; computed on the same hidden-state pools as CKA in Table 4.

D Probe Selectivity

Table 9 reports linear-probe accuracy, MLP-probe accuracy, shuffled-label control accuracy, selectivity (*linear* – *control*), and MLP–linear gap at each

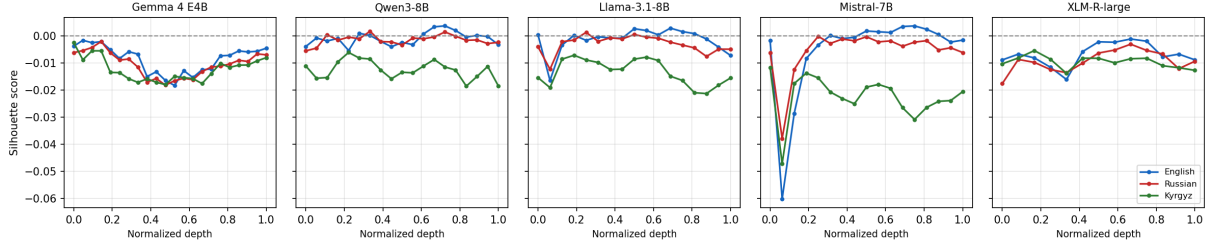


Figure 1: Silhouette score per layer for all five models (cosine metric, 500 subsampled points; x -axis normalized to $[0, 1]$ for cross-model comparison). Values are near zero across all (model, language, layer) cells, confirming that emotion is linearly decodable in these models without forming compact geometric clusters. The pattern holds uniformly across encoder (XLM-R) and decoder architectures.



Figure 2: UMAP projections of hidden states for Gemma 4 E4B (top) and Qwen3-8B (bottom), the two models with the most contrasting depth profiles, at their unified best probing layers (Table 3: Gemma EN L0/RU L7/KY L4; Qwen3 EN L10/RU L11/KY L9). Each point is one sentence, colored by emotion. Both models lack clear geometric clustering, consistent with the near-zero silhouette scores (Figure 1).

model’s best layer. Selectivity is positive and substantial for every cell (0.29–0.46), confirming that probes decode genuine emotion structure rather than representational artifacts. MLP gains over linear probes are small (mean +0.016 across 12 decoder cells) and not tested for significance.

E Sentence-Length Control

Table 10 reports English probing accuracy at each decoder’s best layer, with per-example 5-fold CV predictions binned by source-sentence

word count. Accuracy decreases with length for all four decoders—monotonically across bins for Gemma 4 and Llama, while Qwen3 and Mistral show a small uptick in the final bin—consistent with mean-pooling dilution over longer sequences (Section 6.5).

F Lexical (TF-IDF) Transfer Baseline

Table 11 runs a TF-IDF baseline under the exact leakage-free transfer protocol (same seed-42 60/20/20 split; vectorizer fit on the source lan-

Model	Lang	Linear	MLP	Ctrl	Sel.	Δ MLP
Gemma 4	EN	.603	.621	.191	.413	+.018
Gemma 4	RU	.508	.534	.164	.345	+.026
Gemma 4	KY	.453	.485	.162	.292	+.032
Qwen3	EN	.605	.611	.165	.441	+.005
Qwen3	RU	.582	.607	.172	.410	+.025
Qwen3	KY	.482	.506	.183	.299	+.024
Llama	EN	.625	.636	.165	.460	+.011
Llama	RU	.568	.589	.172	.396	+.021
Llama	KY	.478	.482	.169	.309	+.004
Mistral	EN	.611	.619	.152	.459	+.008
Mistral	RU	.570	.581	.176	.393	+.011
Mistral	KY	.451	.455	.149	.301	+.005

Table 9: Probe validity controls at the best layer (decoder LLMs). XLM-R selectivity follows the same qualitative pattern (0.29–0.37; control 0.17–0.18, near $1/6 \approx 0.167$). Sel. = Linear–Ctrl; Δ MLP = MLP–Linear; both are computed on unrounded accuracies and may differ from the difference of the displayed rounded columns by $\leq .001$.

Model	1–5	6–10	11–15	16–20	21+
<i>n</i> per bin	242	362	369	266	241
Gemma 4	.649	.627	.615	.556	.531
Qwen3	.653	.635	.631	.534	.577
Llama	.690	.638	.621	.609	.577
Mistral	.715	.627	.626	.545	.548

Table 10: English probing accuracy by sentence length (words). Per-example correctness from 5-fold CV at each model’s best EN layer, binned by word count of the source sentence.

guage, target test sentences transformed with that vocabulary). It shares the word 1–2-gram featurization with the within-language baseline of Section 6.5 but uses default classifier settings; matching the tuned within-language configuration (min_df=2, sublinear tf, $C=10$, class weighting) *lowers* the cross-lingual bar on three of the four KY directions (KY→EN .132 vs. .253; KY→RU .179 vs. .216; RU→KY .216 vs. .223) and raises EN→KY by only +.003, so the reported bar is the more conservative of the two overall. The operative axis is the *source language*, not the script: with an EN source the baseline is at or below chance (.128–.132; only 3.7–5.4% of target sentences contain any source-vocabulary token), while RU/KY sources reach .196–.253 with vocabulary-hit shares of .30–.69. The largest hit shares are on the *English-target* directions (RU→EN .689, KY→EN .507): the RU/KY translations retain a Latin-script residue (proper names, code-switched English from the Reddit-sourced originals), and RU–KY adds Cyrillic borrowings. That residue also means some English tokens are physically present in the RU/KY inputs—a caveat that touches the logit lens (Appendix O) and the “lexically

language-specific” premise, though at .20–.25 it bounds the effect well below the strong models’ transfer. As the lexical bar for Table 5: the *best model per direction* exceeds it by .14–.26 on every KY direction (on RU→KY only XLM-R clears it by that much; the top decoders there are at +.05–.08); OLMo-2 sits at or below the bar on three of the four (RU→KY –.047, KY→EN –.040, KY→RU –.013), mBERT on two (KY→EN –.067, KY→RU .000; +.003 on RU→KY); both are above it on EN→KY, and Mistral clears it only narrowly outside EN→KY (+.017–.027).

Direction	TF-IDF acc	Vocab-hit share
EN→RU	.132	.037
EN→KY	.128	.054
RU→EN	.196	.689
RU→KY	.223	.304
KY→EN	.253	.507
KY→RU	.216	.405

Table 11: Leakage-free TF-IDF transfer (word 1–2 grams + logistic regression; chance = $1/6 \approx .167$). Vocab-hit share = fraction of target test sentences containing at least one source-vocabulary token.

G Within-language Probing Pairwise Tests

For each language, we run paired bootstrap tests (10,000 resamples on per-example correctness at each model’s unified best layer from Table 3; fixed seed, script provenance in the repository README), with Holm–Bonferroni correction within language (Table 12). Significance patterns here are sensitive to the choice of per-model best layer and to the CV seed (the per-layer curves are flat near their maxima), so we fix both: unified best layers from Table 3 and a fixed seed. No decoder pair survives Holm correction on EN or KY; on RU, Gemma 4 sits significantly below Qwen3, Llama, and Mistral. This matches the pattern read off Table 3’s overlapping CIs. XLM-R and mBERT are not included in this family—encoder–decoder gaps are read from overlapping CIs in Table 3 and from the transfer tests in Appendix H.

H XLM-R vs. Others on KY_{avg}

Paired bootstrap (10,000 resamples) on the 296 held-out test sentences shared across models, on KY_{avg} (per-example mean of the four KY-involving transfer accuracies from Table 5). Holm–Bonferroni correction is applied within the family

Lang	Comparison	Δ_{acc}	p_{Holm}	sig
EN	Llama vs Gemma 4	+0.022	.274	ns
EN	Llama vs Qwen3	+0.020	.760	ns
EN	Llama vs Mistral	+0.014	1.00	ns
EN	Mistral vs Gemma 4	+0.008	1.00	ns
EN	Mistral vs Qwen3	+0.005	1.00	ns
EN	Qwen3 vs Gemma 4	+0.003	1.00	ns
RU	Qwen3 vs Gemma 4	+0.074	<.001	***
RU	Mistral vs Gemma 4	+0.062	<.001	***
RU	Llama vs Gemma 4	+0.060	<.001	***
RU	Qwen3 vs Llama	+0.014	.806	ns
RU	Qwen3 vs Mistral	+0.013	.806	ns
RU	Mistral vs Llama	+0.001	.937	ns
KY	Qwen3 vs Mistral	+0.032	.091	ns
KY	Qwen3 vs Gemma 4	+0.029	.136	ns
KY	Llama vs Mistral	+0.028	.136	ns
KY	Llama vs Gemma 4	+0.025	.179	ns
KY	Qwen3 vs Llama	+0.004	1.00	ns
KY	Gemma 4 vs Mistral	+0.003	1.00	ns

Table 12: Within-language paired-bootstrap tests among the four decoders at their unified best layers, Holm-corrected within each language (6 pairs). $\Delta_{\text{acc}} = \text{model}_{\text{hi}} - \text{model}_{\text{lo}}$ (ordered so $\Delta \geq 0$), computed on raw per-example correctness (may differ by $\leq .001$ from the difference of Table 3 rounded accuracies). *** $p < .001$, ** $p < .01$, * $p < .05$.

of 7 XLM-R-vs-other comparisons.

Model _a	Model _b	Δ	95% CI	p	p_{Holm}
XLM-R	Qwen3	+0.036	[−.004, +0.074]	.081	.081 ns
XLM-R	Llama	+0.043	[+0.003, +0.084]	.040	.080 ns
XLM-R	Gemma 4	+0.092	[+0.050, +0.133]	<.001	<.001 ***
XLM-R	Mistral	+0.145	[+0.107, +0.183]	<.001	<.001 ***
XLM-R	mBERT	+0.186	[+0.149, +0.224]	<.001	<.001 ***
XLM-R	OLMo-2	+0.201	[+0.164, +0.239]	<.001	<.001 ***
XLM-R	Qwen2.5	+0.116	[+0.074, +0.155]	<.001	<.001 ***

Table 13: XLM-R vs. every other model on KY_{avg} (paired bootstrap, two-sided; Holm-corrected within this family). $\Delta = \text{XLM-R} - \text{other}$, computed on the raw per-example correctness vectors (may differ by $\leq .001$ from the difference of displayed rounded accuracies). *** $p < .001$, ** $p < .01$, * $p < .05$. XLM-R’s absolute-KY lead is significant vs. every model except the two strongest decoders (Qwen3 and Llama, both *n.s.* after correction); vs. the second encoder mBERT the gap is +.186 ($p_{\text{Holm}} < .001$). The lack of significance vs. Qwen3/Llama is why we also report the chance-corrected efficiency η_{cc} (Table 5) as a *secondary* metric (Section 6.3; caveats in Appendix J); it is chance-corrected because a raw KY/EN $\leftrightarrow$ RU ratio is misleading for near-chance models such as mBERT (whose EN $\leftrightarrow$ RU is only .082 above the 1/6 chance floor).

Direction interaction (difference-in-differences). The into-KY concentration of Section 6.3 is tested with a paired bootstrap on the same 296 sentences: $\text{DiD} = (\text{XLM-R} - \text{decoder})_{\text{into-KY}} - (\text{XLM-R} - \text{decoder})_{\text{KY-source}}$. Against Qwen3: +.108, 95% CI [.029, .188], $p = .008$; against Qwen3-fp16 (matched precision): +.074 [−.002, .149],

$p = .058$; against Llama: +.002 [−.078, .079], $p = .97$. XLM-R’s advantage over Llama is direction-uniform; the significant interaction is specifically against the decoder that exceeds XLM-R on KY-source (Qwen3).

I mBERT vs. decoders on absolute KY

Paired bootstrap (10,000 resamples of the 296 held-out test sentences) for mBERT vs. each decoder on KY_{avg} , with Holm-Bonferroni over 6 comparisons.

Model _a	Model _b	Δ	95% CI	p	p_{Holm}
mBERT	Qwen3	−.150	[−.189, −.111]	<.001	<.001 ***
mBERT	Llama	−.143	[−.182, −.104]	<.001	<.001 ***
mBERT	Gemma 4	−.094	[−.130, −.058]	<.001	<.001 ***
mBERT	Qwen2.5	−.070	[−.104, −.036]	<.001	<.001 ***
mBERT	Mistral	−.041	[−.072, −.009]	.014	.027 *
mBERT	OLMo-2	+0.015	[−.018, +0.048]	.378	.378 ns

Table 14: mBERT vs. every decoder on absolute KY_{avg} (paired bootstrap, two-sided; Holm-corrected within this 6-comparison family). The second multilingual encoder is significantly below five of six decoders and statistically tied with OLMo-2 at the bottom, confirming that encoder architecture alone does not confer KY transfer (this is the negative claim that carries Section 7 after the η_{cc} CI showed the XLM-R-vs-mBERT gap under η_{cc} is imprecisely estimated).

J Bootstrap CI for η_{cc}

Paired bootstrap (10,000 resamples of the 296 held-out test sentences, jointly recomputing η_{cc} ’s numerator and denominator on the same resample) for the chance-corrected efficiency across all 8 models, and the differences vs. XLM-R. The same procedure applied to the size-control contrast (Section 6.3) gives XLM-R-base−mBERT: $\text{KY}_{\text{avg}} +.105$ [+0.068, +.141], $p < 10^{-4}$; $\eta_{\text{cc}} +.277$ [−.204, +.603], $p = .20$ —the data-swap effect is established on absolute KY but not resolvable under η , whose mBERT denominator sits near chance (base_vs_mbert_eta.csv).

Limits of ratio normalizations. A ratio metric rewards a small denominator, and XLM-R sits at the low end of *every* English-anchored denominator available here: EN $\leftrightarrow$ RU .473 (main decoders .505–.563), EN $\rightarrow$ RU .470, within-language EN probing .551. Swapping among these normalizations therefore tests robustness to the choice *within the family*, not the family’s shared weakness. Within the family the lead is stable: normalizing by EN $\rightarrow$ RU alone (same 296 examples, paired bootstrap) gives XLM-R .718 [.60, .88] vs. Llama

.471 and Qwen3 .440 ($\Delta=+.247$, $p=.002$; $+.279$, $p<.001$; mBERT unresolved [.10, 1.07]); normalizing by EN probing points the same way (.567 vs. Llama .381, Qwen3 .415; mBERT .101) but mixes samples (1,480-sentence 5-fold CV vs. 296 held-out), so we report it without a CI. What breaks the potential circularity is the *ratio-free directional* evidence of Section 6.3: XLM-R’s advantage lives in the into-KY directions and, against Qwen3, disappears on KY-source ones (against Llama it is direction-uniform)—patterns no denominator can induce. η_{cc} also reorders Llama (.487) and Qwen3 (.460) relative to absolute KY (.341 vs. .349), so we do not read decoder–decoder order from any of these ratios.

Model	η_{cc}	95% CI
XLM-R	.710	[+.61, +.83]
Llama	.487	[+.40, +.58]
Qwen3	.460	[+.39, +.53]
mBERT	.390	[+.11, +.85]
Gemma 4	.371	[+.28, +.47]
Qwen2.5	.278	[+.21, +.35]
Mistral	.198	[+.13, +.27]
OLMo-2	.054	[−.03, +.13]

<i>XLM-R vs. others (paired bootstrap on $\Delta\eta_{cc}$):</i>		
XLM-R – Qwen3	+.250	[+.13, +.39] $p<.001$ ***
XLM-R – Llama	+.224	[+.08, +.37] $p=.002$ **
XLM-R – mBERT	+.321	[−.14, +.61] $p=.13$ ns
XLM-R – Gemma 4	+.339	[+.19, +.49] $p<.001$ ***
XLM-R – Qwen2.5	+.433	[+.31, +.56] $p<.001$ ***
XLM-R – Mistral	+.513	[+.40, +.64] $p<.001$ ***
XLM-R – OLMo-2	+.657	[+.53, +.79] $p<.001$ ***

Table 15: Bootstrap CIs for chance-corrected efficiency $\eta_{cc} = (KY_{avg} - 1/6) / (EN \leftrightarrow RU - 1/6)$ (top block) and paired differences vs. XLM-R (bottom block). XLM-R significantly leads all decoders including the strongest (Llama, Qwen3); the XLM-R vs. mBERT difference is not significant because mBERT’s CI is wide—mBERT’s $EN \leftrightarrow RU$ is only .082 above chance, so the ratio is estimated imprecisely. The robust story is that η_{cc} makes XLM-R a significant outlier over every decoder, but mBERT’s position is imprecisely estimated at this sample size and its relative ranking vs. Llama/Qwen3 or XLM-R is not resolvable. The negative claim (“encoder/decoder does not order the 8 models”) is carried instead by absolute KY: mBERT is second-to-last there and significantly below 5 of 6 decoders (Appendix I). η_{cc} is computed from unrounded accuracies; recomputing it from the rounded entries of Tables 5/22 can differ by up to .007 (mBERT: .383 vs. .390).

K Transfer Pairwise Tests

Table 16 reports all six pairwise *leakage-free* transfer-accuracy comparisons among decoder LLMs per direction (paired bootstrap on per-example correctness over the held-out target test split, Holm-Bonferroni corrected within each di-

rection). After removing the parallel-sentence leak, most decoder–decoder differences are no longer significant. The surviving effects concentrate on KY-involving directions: Qwen3 exceeds Mistral on $EN \rightarrow KY$, $KY \rightarrow EN$, and $KY \rightarrow RU$; Llama exceeds Mistral on $EN \rightarrow KY$ and $KY \rightarrow RU$ ($KY \rightarrow EN$ is n.s.); Llama exceeds Gemma 4 on $EN \rightarrow KY$, Qwen3 exceeds Gemma 4 on $KY \rightarrow EN$, and Qwen3 exceeds Llama on $KY \rightarrow EN$ ($+.101$, $p_{Holm}=.025$). $RU \rightarrow KY$ and $EN \leftrightarrow RU$ have no surviving decoder–decoder differences.

L Translation-source Control

Table 17 reports, for each model and language (RU, KY), the probing accuracy on the 102-sentence control set (Section 3.3) with (a) Claude-translated and (b) human-translated versions of the same source EN sentences. A linear probe is trained on the remaining 1,378 Claude-translated examples at the model’s best layer (per Table 3) and applied to both versions of the held-out 102. The 95% CI is computed over 10,000 paired bootstrap resamples of the 102 predictions.

Transfer under human translations. The confound is most threatening for *transfer* (a shared generator could inflate cross-lingual alignment), so we repeat the check there: $EN \rightarrow RU$ and $EN \rightarrow KY$ probes—trained on the Claude training split excluding the 102 control sentences, at the transfer layers of Table 5—are evaluated on the Claude- vs. human-translated target versions of the 102. All ten (model, direction) cells are n.s. ($|\Delta| \leq .029$, signs in both directions, paired bootstrap $p \geq .26$). At $n=102$ the CIs are wide, but there is no directional evidence that Claude-translated targets are easier to transfer into.

M Pooling Ablation: Mean vs. Last-Token

We re-extract hidden states from all four decoder LLMs using *last-token* pooling (the hidden state at the final non-padding token, located robustly via the attention mask) and re-run layer-wise probing under the same protocol as Section 6. XLM-R is omitted because last-token pooling is not meaningful for a bidirectional encoder. Table 18 reports the best layer and accuracy under each pooling choice.

A natural interpretation: mean pooling integrates emotion-bearing lexical signal already present at the embedding layer (words like *furious*, *delighted*),

Direction	Comparison	Δacc	p_{Holm}	sig	Direction	Comparison	Δacc	p_{Holm}	sig
EN→RU	Qwen3 vs Gemma 4	+0.084	.092	ns	RU→KY	Llama vs Mistral	+0.064	.168	ns
EN→RU	Qwen3 vs Llama	+0.044	.737	ns	RU→KY	Gemma 4 vs Mistral	+0.051	.896	ns
EN→RU	Qwen3 vs Mistral	+0.044	.737	ns	RU→KY	Qwen3 vs Mistral	+0.037	1.00	ns
EN→RU	Llama vs Gemma 4	+0.041	.737	ns	RU→KY	Llama vs Qwen3	+0.027	1.00	ns
EN→RU	Mistral vs Gemma 4	+0.041	.737	ns	RU→KY	Gemma 4 vs Qwen3	+0.014	1.00	ns
EN→RU	Llama vs Mistral	+0.000	1.00	ns	RU→KY	Llama vs Gemma 4	+0.014	1.00	ns
EN→KY	Llama vs Mistral	+0.159	<.001	***	KY→EN	Qwen3 vs Mistral	+0.162	<.001	***
EN→KY	Llama vs Gemma 4	+0.108	.002	**	KY→EN	Qwen3 vs Gemma 4	+0.122	.004	**
EN→KY	Qwen3 vs Mistral	+0.095	.045	*	KY→EN	Qwen3 vs Llama	+0.101	.025	*
EN→KY	Llama vs Qwen3	+0.064	.107	ns	KY→EN	Llama vs Mistral	+0.061	.205	ns
EN→KY	Gemma 4 vs Mistral	+0.051	.299	ns	KY→EN	Gemma 4 vs Mistral	+0.041	.523	ns
EN→KY	Qwen3 vs Gemma 4	+0.044	.299	ns	KY→EN	Llama vs Gemma 4	+0.020	.585	ns
RU→EN	Qwen3 vs Gemma 4	+0.030	1.00	ns	KY→RU	Qwen3 vs Mistral	+0.145	<.001	***
RU→EN	Qwen3 vs Llama	+0.030	1.00	ns	KY→RU	Llama vs Mistral	+0.125	.002	**
RU→EN	Qwen3 vs Mistral	+0.017	1.00	ns	KY→RU	Qwen3 vs Gemma 4	+0.074	.197	ns
RU→EN	Mistral vs Gemma 4	+0.014	1.00	ns	KY→RU	Gemma 4 vs Mistral	+0.071	.197	ns
RU→EN	Mistral vs Llama	+0.014	1.00	ns	KY→RU	Llama vs Gemma 4	+0.054	.344	ns
RU→EN	Gemma 4 vs Llama	+0.000	1.00	ns	KY→RU	Qwen3 vs Llama	+0.020	.561	ns

Table 16: Leakage-free pairwise transfer comparisons among decoder LLMs ($\Delta\text{acc} = \text{model}_{\text{hi}} - \text{model}_{\text{lo}}$, ordered so $\Delta \geq 0$; computed on raw per-example correctness, may differ by $\leq .001$ from the difference of displayed rounded accuracies). p -values from a paired bootstrap (10,000 resamples) on per-example correctness over the held-out target test split, Holm–Bonferroni corrected within each direction (6 pairs). *** $p < .001$, ** $p < .01$, * $p < .05$. Most differences are n.s. after removing the parallel-sentence leak; surviving effects are on KY-involving directions only—see main text for the exact set of significant comparisons.

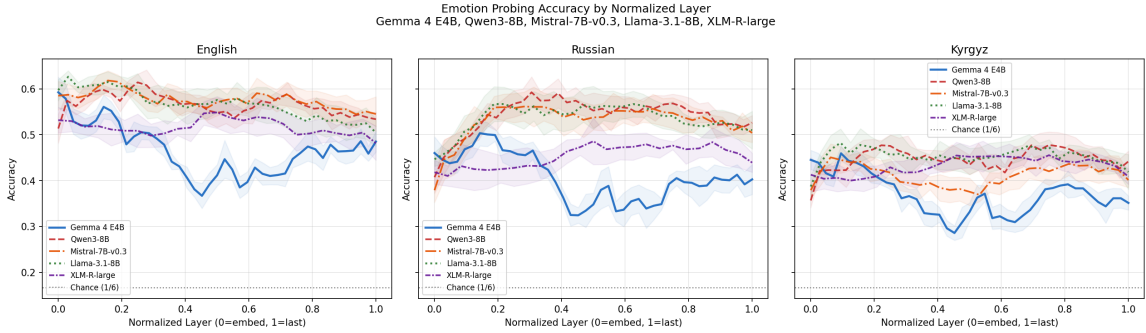


Figure 3: Layer-wise probing accuracy for the four decoder LLMs and XLM-R across three languages. The x -axis is normalized to $[0, 1]$ for cross-model comparison (note: this equates architecturally non-equivalent positions across models of different depths). Shaded bands show ± 1 std (5-fold CV). A depth trend is visible on English as a continuous spread in normalized depth (Gemma 4/Llama at $L/N \leq .03$, Mistral .16, Qwen3 .28, XLM-R .46) but does not hold uniformly across languages—e.g., Llama RU peaks at L19 (Table 3), deeper than Qwen3 RU.

so probing succeeds early. Last-token pooling requires the model to *summarize* that signal into a single position—a function not explicitly trained for on unprompted text—and thus needs more layers of contextual processing. Both metrics are valid; we report mean pooling as primary because it yields the more decodable representation across all conditions.

N Quantization Ablations

Gemma 4: NF4 vs. 8-bit probing. To bound the effect of 4-bit NF4 quantization on probing accuracy, we re-extracted Gemma 4 hidden states in 8-bit (LLM.int8) and re-ran layer-wise probing under the same protocol (Table 19). The accuracy difference is at most .016 in absolute value and

changes sign across languages, i.e. quantization neither systematically helps nor hurts at the resolution of our experiment. Both columns come from the ablation’s own matched probing run, whose NF4 accuracies differ by $\leq .010$ from the unified run of Table 3 (separate CV seed).

Qwen3: NF4 vs. fp16 transfer (matched precision with the encoders). We re-extracted Qwen3-8B end-to-end in fp16 and re-ran the full leakage-free transfer protocol (same seed-42 splits, nested dev layer selection; Table 20). Individual directions shift by up to $\pm .09$ as the nested selection picks different layers on flat dev curves, but KY_{avg} moves only from .349 to .369 ($\Delta = +.020$, paired bootstrap 95% CI $[-.007, +.047]$, $p = .16$), and

Model	Lang	L	Acc _C	Acc _H	Δ	95% CI
Gemma 4	RU	7	.539	.539	.000	[.000, .000]
Gemma 4	KY	4	.480	.451	+.029	[−.049, +.108]
Qwen3	RU	11	.578	.578	.000	[.000, .000]
Qwen3	KY	9	.520	.480	+.039	[−.039, +.118]
Llama	RU	19	.500	.490	+.010	[.000, +.029]
Llama	KY	3	.451	.431	+.020	[−.059, +.098]
Mistral	RU	9	.588	.588	.000	[.000, .000]
Mistral	KY	5	.441	.422	+.020	[−.049, +.088]
XLm-R	RU	21	.490	.490	.000	[.000, .000]
XLm-R	KY	14	.500	.441	+.059	[−.020, +.137]

Table 17: Probing accuracy on the 102-sentence control set with Claude-translated (C) vs. human-translated (H) versions of the same source EN sentences. Probes are trained on the remaining 1,378 Claude-translated examples at the best layer per (model, language) and applied to both versions of the held-out 102. $\Delta = \text{Acc}_C - \text{Acc}_H$; 95% CIs from 10,000 paired bootstrap resamples. All 10 CIs include zero (one at the boundary: Llama RU [.000, +.029]); the maximum observed difference is 5.9 pp (XLm-R KY, n.s.). For several models on Russian $\Delta = 0$ exactly: this is not a data error—the human and Claude translations differ in *all* 102 sentences (0 textually identical) and produce distinct hidden states (max $|\Delta h| = 4.5$ at the probed layer)—rather, the probe assigns the *same emotion label* to both versions of each test sentence, i.e. its decisions are invariant to the translation source even though the inputs are not. A prediction-distribution audit (an independent probe re-run; cell accuracies within .03 of those shown) confirms the probes are not degenerate: in every cell all six classes are predicted at near-uniform frequencies (entropy ≥ 1.74 of the $\ln 6 \approx 1.79$ maximum), and the C/H predictions agree for 100–102 of 102 RU sentences vs. 74–79 on KY (audit table in the repository). We read this as robustness of the probe, not as identity of the representations.

the XLm-R – Qwen3 gap at matched precision is +.015 ([−.025, +.054], $p=.46$)—still positive, still not significant. The absolute-KY reading of Section 6.3 is therefore not an artifact of comparing NF4 decoders with fp16 encoders.

O Logit-Lens Test of the English Pivot

To test whether high EN $\leftrightarrow$ RU transfer reflects genuinely language-agnostic representations or an English pivot (Wendler et al., 2024), we apply a sentence-level *logit lens* to the three dense decoders: Qwen3-8B (highest mean EN $\leftrightarrow$ RU transfer, .56), Llama-3.1-8B (peak CKA and Procrustes similarity across EN–RU; Table 4, Appendix C), and Mistral-7B (the CKA–transfer dissociation case). Gemma 4 is excluded: its Per-Layer Embeddings add per-layer input terms, so a pooled hidden

Model	Lang	Mean pooling			Last-token		
		L	L/N	Acc	L	L/N	Acc
Gemma 4	EN	0	.00	.603	42	1.00	.383
Gemma 4	RU	7	.17	.508	7	.17	.348
Gemma 4	KY	4	.10	.453	7	.17	.312
Qwen3	EN	10	.28	.605	36	1.00	.476
Qwen3	RU	11	.31	.582	35	.97	.447
Qwen3	KY	9	.25	.482	11	.31	.374
Llama	EN	1	.03	.625	28	.88	.472
Llama	RU	19	.59	.568	14	.44	.474
Llama	KY	3	.09	.478	16	.50	.397
Mistral	EN	5	.16	.611	10	.31	.473
Mistral	RU	9	.28	.570	17	.53	.455
Mistral	KY	5	.16	.451	32	1.00	.336

Table 18: Best probing layer under mean vs. last-token pooling. L = absolute best layer; L/N = normalized to network depth. Mean pooling yields higher accuracy in every cell (drops .081–.220 under last-token); last-token pooling shifts best layers substantially later on English for three of four models (Gemma 0 $\rightarrow$ 42, Qwen3 10 $\rightarrow$ 36, Llama 1 $\rightarrow$ 28), indicating that the early-peaking pattern is pooling-dependent (Section 7). Mean-pooling numbers match the unified probing run (Table 3); last-token values from separate extraction on the same corpus.

Lang	L _{NF4}	Acc _{NF4}	L _{8bit}	Acc _{8bit}	Δ
EN	0	.593	0	.599	+.006
RU	6	.503	6	.519	+.016
KY	4	.459	5	.445	−.015

Table 19: Gemma 4 best-layer probing accuracy under 4-bit NF4 vs. 8-bit (LLM.int8) extraction. $\Delta = 8\text{-bit} - \text{NF4}$, computed on unrounded accuracies (may differ from the difference of the rounded columns by $\leq .001$); best layers agree (EN, RU) or are adjacent (KY).

Direction	L _{NF4} [*]	Acc _{NF4}	L _{fp16} [*]	Acc _{fp16}
EN $\rightarrow$ RU	15	.581	22	.581
EN $\rightarrow$ KY	12	.294	17	.385
RU $\rightarrow$ EN	10	.544	11	.551
RU $\rightarrow$ KY	13	.277	31	.260
KY $\rightarrow$ EN	20	.443	20	.402
KY $\rightarrow$ RU	19	.382	14	.429
KY _{avg}		.349		.369

Table 20: Qwen3-8B leakage-free transfer under NF4 vs. fp16 extraction (nested layer selection re-run per precision; L^{*} = selected layer). KY_{avg}: $\Delta = +.020$, n.s.; XLm-R–Qwen3(fp16) = +.015, n.s.

state does not share the `lm_head` input space the lens assumes. For each layer we take the mean-pooled hidden state (the same representation used for probing), apply the model’s final RMSNorm, and project it through the unembedding matrix (`lm_head`) to obtain a distribution over the vocabulary. We decode the top-20 tokens and classify each as English (Latin-script word token), Cyrillic (input-language family for RU/KY), or other

(punctuation, digits, sub-word byte fragments, or non-target scripts). We report, per layer and averaged over 300 sentences, the fraction of top-20 tokens in each class.

The main-text result and its interpretation are in Section 6.4; Figure 4 gives the full per-layer curves. We add here the two model-level contrasts visible in the figure: (i) Llama and Mistral exhibit cleaner, wider English basins than Qwen3, whose early-mid layers carry a larger “other” (code/multilingual) fraction; (ii) Kyrgyz leaves the basin in the last third of depth in all three models ($d_{.5} \approx .97/.86/.97$) and never before Russian, whose exit depth is model-specific ($d_{.5} \approx .97/.64/.75$); the *manner* of exit, by contrast, varies by model, as the AUC and text-specified-window analyses below show (Section 6.4). As a pooled, sentence-level lens it speaks to the aggregate quantity relevant to our pooled-probing transfer; a per-token lens (Wendler et al., 2024) would localize *where* in the sequence the pivot occurs and is left for future work.

Shuffled-representation null. To calibrate the unembedding’s frequency bias we run the null directly: the same 300 pooled vectors per layer, each with its dimensions randomly permuted (destroying learned structure while preserving the vector’s norm and coordinate statistics), passed through the same final RMSNorm and `lm_head` and classified identically. The null decodes to English almost exclusively and *flat in depth*: the *renormalized* word-like EN share, $\text{EN}/(\text{EN}+\text{Cyrillic})$, is .952–.967 (Llama), .967–.981 (Qwen3), and .963–.977 (Mistral) at *every* layer, while the *raw* Cyrillic fraction of the top-20 never exceeds .042. Two conclusions follow. First, the unembedding’s high-logit region is strongly English-biased, so the absolute mid-network English level in Figure 4 is not by itself evidence of English-like computation—indeed, the real curves’ mid-network means (renormalized EN share over $[-.2, .6]$ depth) are .99 (Llama RU), .95 (Llama KY), .94 (Qwen3 KY), .77 (Qwen3 RU), .97 (Mistral RU), and .99 (Mistral KY), i.e. at or below the null except marginal excesses for Llama-RU and Mistral-KY. Qwen3-RU’s .20 *deficit* below the null is itself informative: decoding *less* to English than a shuffled vector requires structured non-English content in the representation—again structure, not unembedding bias.

Second, everything the null cannot produce is informative: the real curves’ late-layer Cyrillic

take-over (up to 1.0), its abrupt-vs-gradual manner, and its language- and model-specific depth (Section 6.4) all require the *representations* to change across depth—the English-pivot claim rests on this structured exit, not on the raw English share.

Sentence-level uncertainty. The $\text{AUC}_{[.5,1]}$ contrasts carry the per-model claims, so we bootstrap them over the 300 sentences (trapezoid AUC is linear in the mean curve, so the mean of per-sentence AUCs equals the AUC of the mean curve; 10,000 resamples). All three RU-vs-KY contrasts are tight and significant: $\Delta\text{AUC}(\text{KY}-\text{RU}) = -.162 [-.170, -.153]$ for Llama (Russian holds the basin *longer*), $+.380 [.367, .393]$ for Qwen3, and $+.264 [.247, .280]$ for Mistral (all $p < 10^{-4}$). Sentences whose top-20 contains no word-like token are imputed share 1.0—counted as fully English, which biases toward *later* exit and more strongly for KY (more word-free decodes), the direction of our conclusion; the aggregate effect is bounded: pooled-count curves and per-sentence means differ by $\leq .003$ (e.g., Qwen3-RU AUC .236 vs. .238). AUC measures how much English is *held* over the second half—exit level and timing, not abruptness—and the per-model signs differ exactly as described in Section 6.4. For Mistral’s quoted window (.56–.84) the per-sentence paired bootstrap gives an RU–KY drop difference of $+.557 [.529, .584]$, $p < 10^{-4}$; since that window was chosen on the curve it tests, we treat it as descriptive. A fixed window *specified by the text* (the last third, $[\cdot67, 1]$; fixed before this test but not registered in advance) applied identically to all three models reduces to the level at .67, because every model’s final layer decodes the input language ($\text{share}(1.0) \leq .005$): RU–KY drop differences are $+.085$ (Llama), $-.636$ (Qwen3), $-.196$ (Mistral), all tightly estimated, reproducing the same per-model ordering as the AUC gaps (note the mirrored sign convention: RU–KY drops vs. KY–RU AUC) rather than a uniform manner contrast. We therefore present “gradual vs. steep” as per-model description (quantified by each model’s $d_{.5}/\text{AUC}$ and, for Mistral, the descriptive window), not as a single cross-model statistic.

Robustness: classification and pooling. The lens has two obvious vulnerabilities, and the exit result survives both (`lens_robustness.csv`). *Classification*: since KY fertility exceeds RU’s, Cyrillic could be undercounted more for KY, pushing $d_{.5}$ later—the direction of our conclusion. Re-classifying any token containing a Cyrillic charac-

ter as Cyrillic changes nothing (mixed tokens are rare in the top-20); additionally attributing every undecodable or letterless fragment to Cyrillic—an upper bound—moves $d_{.5}$ by at most .03 (Qwen3 KY .86→.83, Mistral KY .97→.94) and preserves $KY \geq RU$ in all three models, with all AUC signs intact. *Pooling*: the mean-pooled vector is not a residual-stream state. A last-token lens on the same 300 sentences yields $d_{.5}$ *ties* at .92–.97 for both languages in all three models—at the final token the input language re-emerges only at the very end, consistent with per-token analyses—while the AUC signs match the pooled lens (Llama RU>KY; Qwen3/Mistral KY>RU). The strict pooled Russian exits at .64 (Qwen3) and .75 (Mistral) are thus a property of the pooled aggregate; the ordering claim (KY never earlier, always last-third) holds under both poolings. *Sampling*: a sentence-level bootstrap (2,000 resamples of the 300 sentences; the mean curve and sustained crossing recomputed per resample) leaves every $d_{.5}$ at its point estimate—the 95% CI has zero width at the layer grid (1/32; 1/36 for Qwen3)—and $KY \geq RU$ holds in 100% of resamples for each model (d50_bootstrap.csv).

Note that Kyrgyz tokenizer fertility is $2.88\text{--}3.07\times$ English for the three lens models ($1.42\text{--}3.67\times$ across all 8; Table 24), so a top- k decode of a KY-input hidden state contains *more* subword fragments than the English or Cyrillic word-like tokens counted here—i.e., the Cyrillic fraction is a lower bound and the gradual re-emergence of Kyrgyz Cyrillic is measured conservatively; if anything, fragmentation biases the count toward English.

P Expanded Model Sample (N=8)

To strengthen claims that originally rested on five models, we add three: **mBERT** (Devlin et al., 2019) (bert-base-multilingual-cased; 178M, 12 layers; a second multilingual encoder pre-trained, with MLM like XLM-R (plus next-sentence prediction), on the 104 largest Wikipedias, two orders of magnitude less data than XLM-R; Kyrgyz is included in the 104-language set, so the data-composition claim we make is about *scale* rather than about presence/absence of the language), **Qwen2.5-7B** (Qwen Team, 2024) (the previous Qwen generation, for within-family stability with Qwen3-8B), and **OLMo-2-7B** (OLMo Team, 2024) (a decoder LLM with fully disclosed, predominantly English pre-training data; included to probe whether pre-training-data multilingual-

ity rather than encoder architecture drives cross-lingual transfer). As a size control we additionally run **XLM-R-base** (xlm-roberta-base; 270M, 12 layers; same CC-100 corpus and MLM objective as XLM-R-large) through the identical pipeline (fp16 extraction, mean pooling, seed-42 splits, target-dev layer selection); it appears as a control row in Table 22 and is excluded from the core-8 rankings. Its per-layer probing peaks are EN .494 (L0), RU .421 (L11), KY .410 (L3)—below XLM-R-large’s, consistent with a depth/size contribution to absolute accuracy.

Model	EN			RU			KY		
	L	L/N	Acc	L	L/N	Acc	L	L/N	Acc
Gemma 4 E4B	0	.00	.603	7	.17	.508	4	.10	.453
Qwen3-8B	10	.28	.605	11	.31	.582	9	.25	.482
Llama-3.1-8B	1	.03	.625	19	.59	.568	3	.09	.478
Mistral-7B	5	.16	.611	9	.28	.570	5	.16	.451
XLM-R-large	11	.46	.551	21	.88	.482	14	.58	.459
<i>Added in this appendix:</i>									
mBERT	1	.08	.482	2	.17	.374	3	.25	.402
Qwen2.5-7B	9	.32	.588	8	.29	.589	3	.11	.453
OLMo-2-7B	0	.00	.625	18	.56	.534	5	.16	.451
<i>Size control (excluded from core-8 rankings):</i>									
XLM-R-base	0	.00	.494	11	.92	.421	3	.25	.410

Table 21: Best probing layer and accuracy across the expanded 8-model sample plus an XLM-R-base size control. Best EN layers span L/N .00–.46 continuously, with no clean early/late split (four models at $L/N \leq .10$: Gemma 4, Llama, OLMo-2, mBERT; Qwen2.5 and Qwen3 both mid-network, showing within-family stability), consistent with the graded depth shift of Section 6.1.

Q Expanded-Sample CKA

Table 23 reports peak linear CKA for the four models not covered by Table 4, computed with the same procedure on the same hidden-state pools; Section 6.2 gives the reading (XLM-R’s unremarkable CKA vs. its leading KY transfer reinforces the CKA–transfer dissociation).

R Tokenizer Fertility

Table 24 reports subword fertility (subwords per whitespace word) on our corpus for all 8 tokenizers. Fertility enters the Discussion as a weak, sign-consistent but unreliable correlate of KY transfer, and motivates the conservative reading of the logit-lens Cyrillic fractions (Appendix O).

S Emotion Confusions

Per-emotion macro-F1 (Table 6) shows how separable each category is, not which categories are mis-

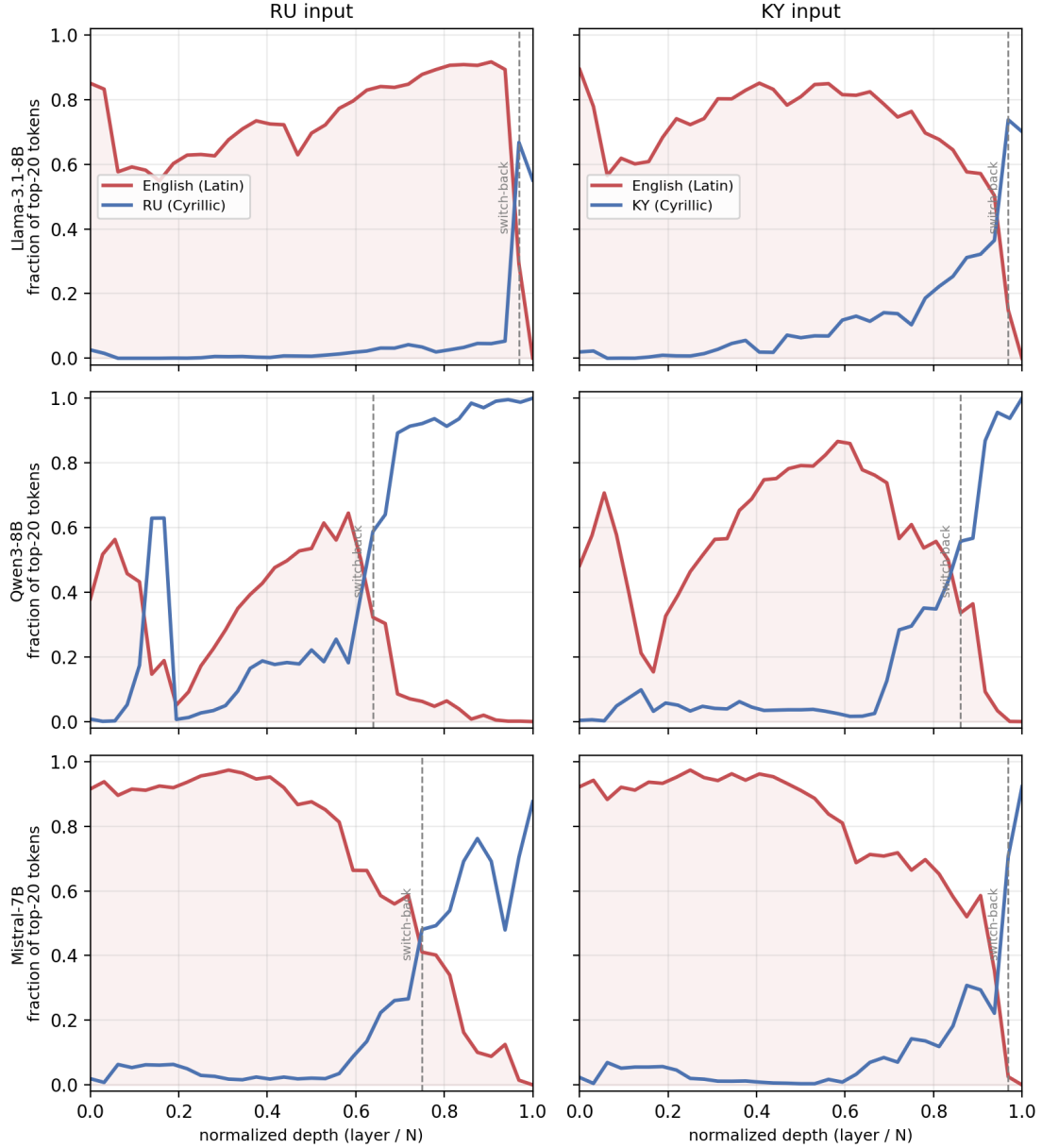


Figure 4: Sentence-level logit lens for Llama-3.1-8B (top), Qwen3-8B (middle), and Mistral-7B (bottom) on Russian (left) and Kyrgyz (right) inputs. Each curve is the *raw* fraction of the top-20 unembedding-projected tokens that are English (Latin, red) vs. input-language Cyrillic (blue), averaged over 300 sentences, as a function of normalized network depth (the remainder to 1.0 is the “other” class: punctuation, digits, sub-word fragments, non-target scripts). The dashed line marks the switch-back: the first layer from which the Cyrillic fraction exceeds the English fraction through to the output. English dominates the mid layers in all six panels—a level the English-biased unembedding reproduces even for shuffled vectors (see the null in Appendix O)—but the exit differs: Kyrgyz switches back only in the last third of depth in all three models ($d_{.5} \approx .97/.86/.97$) and never before Russian, whose switch-back depth is model-specific ($d_{.5} \approx .97/.64/.75$). Section 6.4 quantifies this with $d_{.5}$ and $AUC_{[.5,1]}$ on the word-like-renormalized English share.

taken for one another. We therefore tabulate the six-way confusion matrix of the linear probe for each of the 15 main (model, language) cells, from the per-example 5-fold CV predictions at the unified best layers (re-running the probe of Table 3 with the same seed reproduces every accuracy to four deci-

imals). For a pair of emotions (a, b) we report the symmetric confusion rate $(C_{ab} + C_{ba}) / (n_a + n_b)$: the share of a -or- b sentences assigned to the other member of the pair (Table 25).

Three patterns hold. (i) *Anger is the hub*: Anger–Disgust, Anger–Surprise and Anger–Sadness are

Model	E→R	E→K	R→E	R→K	K→E	K→R	KY _{avg}
Gemma 4	.497	.250	.513	.290	.321	.307	.292
Qwen3	.581	.294	.544	.277	.443	.382	.349
Llama	.537	.358	.513	.304	.341	.361	.341
Mistral	.537	.199	.527	.240	.280	.236	.239
XLm-R	.470	.392	.476	.358	.409	.378	.384
XLm-R-base	.375	.297	.368	.284	.321	.311	.303
mBERT	.250	.166	.247	.226	.186	.216	.198
Qwen2.5	.564	.233	.503	.274	.301	.267	.269
OLMo-2	.436	.142	.517	.176	.213	.203	.183

Table 22: Leakage-free cross-lingual transfer accuracy across the 8 core models plus an XLm-R-base size control (source-trained probes; disjoint sentence splits, nested target-dev layer selection). Encoder rankings quoted in the text refer to the 8 core models. Column headers abbreviate languages: E = English, R = Russian, K = Kyrgyz; KY_{avg} averages the four KY-involving directions. Bold = best per column. XLm-R attains the highest KY transfer (.384), but the ordering is *not* a clean encoder–decoder split: mBERT, also an encoder, ranks seventh of eight (.198, above only OLMo-2 .183), and five decoders (Qwen3, Llama, Gemma 4, Qwen2.5, Mistral) fall between the two encoders. The XLm-R–vs–mBERT gap (.384 vs .198) tracks pre-training data scale and domain (~2.5 TB CommonCrawl vs. ~21 GB Wikipedia; language coverage is comparable, 100 vs. 104), not encoder architecture. The chance-corrected efficiency η_{cc} (Table 5; Appendix J) confirms XLm-R as a significant outlier over the best decoders (vs. Llama $p < .01$, vs. Qwen3 $p < .001$); mBERT’s η_{cc} point estimate is .390 but its CI is wide ([.11, .85]) and its ranking vs. Llama/Qwen3 is not resolvable at $n=296$.

Model	EN–RU	EN–KY	RU–KY
XLm-R	.744 (10)	.682 (10)	.682 (10)
mBERT	.700 (6)	.621 (5)	.657 (5)
Qwen2.5	.673 (27)	.648 (27)	.644 (23)
OLMo-2	.608 (21)	.316 (9)	.413 (10)

Table 23: Peak linear CKA (peak layer in parentheses) for the four models not covered by Table 4, on the same hidden-state pools. The values reinforce the CKA–transfer dissociation of Section 6.2: XLm-R’s peaks (.68–.74) sit below Llama’s and Mistral’s on every language pair (Table 4) although XLm-R leads KY transfer, while OLMo-2 is lowest on both KY pairs (EN–KY CKA .316, RU–KY .413) and on KY transfer (.183). XLm-R’s EN–KY and RU–KY peaks coincide only after rounding (.6818 vs. .6819, distinct per-layer curves); the same holds for Qwen3 in Table 4 (.6203 vs. .6201).

the three most confused pairs on the 5-model mean in EN and RU (.14–.17) and on the cross-language mean, matching Anger’s lowest macro-F1; on KY Disgust–Fear (.156) rises to second place among them. Pairs sometimes expected to be close are not: Fear–Surprise is among the least confused (.05–.08). (ii) *The structure is shared across lan-*

Model	fert _{EN}	fert _{RU}	fert _{KY}	KY/EN
XLm-R	1.43	1.79	2.03	1.42
mBERT	1.50	2.19	2.81	1.87
Gemma 4	1.29	1.84	3.28	2.54
Mistral	1.39	2.69	4.01	2.88
Qwen3	1.28	2.29	3.81	2.97
Qwen2.5	1.28	2.29	3.81	2.97
Llama	1.28	2.13	3.92	3.07
OLMo-2	1.28	3.06	4.68	3.67

Table 24: Tokenizer fertility (subwords per whitespace-word) on our corpus, sorted by KY/EN ratio. XLm-R fragments Kyrgyz least ($1.42\times$ EN, close to Turkic-aware subword coverage); OLMo-2 worst ($3.67\times$). mBERT, despite covering 104 languages, fragments KY second-least (2.81) yet markedly worse than XLm-R (2.03), consistent with its two-orders-of-magnitude smaller training corpus. Across the 8 models, KY fertility correlates negatively with KY_{avg} transfer (Pearson $r = -.46$, $p = .25$; Spearman $\rho = -.49$, $p = .22$; *n.s.* at $n=8$). KY/EN is computed on unrounded fertilities and may differ from the ratio of the rounded columns by $\leq .01$.

guages: within each model the ranking of the 15 pairs correlates at Spearman $\rho = .78$ –.97 between languages (15 model–language-pair values, all $p < .001$; lowest for Qwen3 EN–KY). Because the corpus is parallel and labels are inherited from the English source, part of this agreement reflects shared content and label noise rather than the models alone. (iii) *The EN→KY loss is not uniform*: 13 of the 15 pairs become more confused from EN to KY, most of all Disgust–Fear (.094→.156), while Anger–Sadness is flat (.148) and Anger–Disgust falls (.168→.153)—Kyrgyz blurs distinctions that are clean in English rather than deepening the one that is already hardest. Confusion is also graded within negative valence (Anger–Fear .07 vs. Anger–Disgust .16), so the probes recover more than a positive/negative split. All 15 full matrices are in the repository (`confusion_matrices.csv`).

References

- Guillaume Alain and Yoshua Bengio. 2017. Understanding intermediate layers using linear classifier probes. In *ICLR Workshop*.
- Anthropic. 2025. Claude 3.7 Sonnet system card. Technical report, Anthropic.
- Anthropic. 2026. Claude Opus 4.8. Technical report, Anthropic.
- Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Pair	EN	RU	KY
Anger–Disgust	.168	.154	.153
Anger–Surprise	.138	.145	.160
Anger–Sadness	.148	.137	.148
Disgust–Fear	.094	.135	.156
Sadness–Surprise	.097	.102	.134
Joy–Surprise	.092	.102	.109
Disgust–Sadness	.082	.098	.101
Fear–Sadness	.061	.079	.095
Disgust–Surprise	.066	.074	.082
Anger–Fear	.054	.073	.088
Joy–Sadness	.054	.061	.091
Fear–Surprise	.050	.070	.082
Anger–Joy	.045	.060	.087
Disgust–Joy	.034	.047	.062
Fear–Joy	.026	.040	.062

Table 25: Symmetric pairwise confusion rate of the six-way linear probe, averaged over the 5 main models, at the unified best layers (sorted by the cross-language mean). Under uniformly random six-way predictions the rate would be $1/6 \approx .167$ for every pair.

Sven Buechel and Udo Hahn. 2017. EmoBank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *Proceedings of EACL*, pages 578–585.

Nadezhda Chirkova and Vassilina Nikoulina. 2024. Zero-shot cross-lingual transfer in instruction tuning of large language models. In *Proceedings of the 17th International Natural Language Generation Conference (INLG)*, pages 695–708.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of ACL*, pages 8440–8451.

MohammadReza Davari, Stefan Horoi, Amine Natik, Guillaume Lajoie, Guy Wolf, and Eugene Belilovsky. 2023. Reliability of CKA as a similarity measure in deep learning. In *Proceedings of ICLR*.

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 ACL*, pages 4040–4054.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. *Advances in Neural Information Processing Systems*, 36.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Frances Ding, Jean-Stanislas Denain, and Jacob Steinhardt. 2021. Grounding representation similarity

through statistical testing. In *Advances in Neural Information Processing Systems*, volume 34.

Clément Dumas, Veniamin Veselovsky, Giovanni Monea, Robert West, and Chris Wendler. 2024. Separating tongue from thought: Activation patching reveals language-agnostic concept representations in transformers. *arXiv preprint arXiv:2411.08745*.

Paul Ekman. 1992. An argument for basic emotions. *Cognition and Emotion*, 6(3-4):169–200.

Gemma Team, Google DeepMind. 2026. **Gemma 4 technical report**. <https://arxiv.org/abs/2607.02770>. Preprint, arXiv:2607.02770.

Kilem L. Gwet. 2008. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61(1):29–48.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Proceedings of EMNLP-IJCNLP*, pages 2733–2743.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Léo Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. Technical report, Mistral AI. ArXiv:2310.06825; we use the v0.3 checkpoint.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. 2019. Similarity of neural network representations revisited. In *Proceedings of ICML*, pages 3519–3529.

Meta AI. 2024. The Llama 3 herd of models. Technical report, Meta.

Jamshidbek Mirzakhlov, Anoop Babu, Duygu Ataman, Sherzod Kariev, Francis Tyers, Otabek Abdurafof, Mammad Hajili, Sardana Ivanova, Abror Khaytbaev, Antonio Laverghetta Jr., Bekhzodbek Moydinboyev, Esra Onal, Shaxnoza Pulatova, Ahsan Wahab, Orhan Firat, and Sriram Chellappan. 2021. A large-scale study of machine translation in Turkic languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5876–5890.

Saif M. Mohammad. 2018. Word affect intensities. In *Proceedings of LREC*.

- Shamsuddeen Hassan Muhammad, Nadjma Ousidhoum, Idris Abdulmumin, Jan Philip Wahle, Terry Ruas, Meriem Beloucif, Christine de Kock, Nirmal Surange, Daniela Teodorescu, Ibrahim Said Ahmad, David Ifeoluwa Adelani, Alham Fikri Aji, Felermينو D. M. A. Ali, Ilseyar Alimova, Vladimir Araujo, Nikolay Babakov, Naomi Baes, Ana-Maria Bucur, Andiswa Bukula, and 29 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, Vienna, Austria.
- Shamsuddeen Hassan Muhammad, Nadjma Ousidhoum, Idris Abdulmumin, Seid Muhie Yimam, Jan Philip Wahle, Terry Lima Ruas, Meriem Beloucif, Christine De Kock, Tadesse Destaw Belay, Ibrahim Said Ahmad, Nirmal Surange, Daniela Teodorescu, David Ifeoluwa Adelani, Alham Fikri Aji, Felermينو Dario Mario Ali, Vladimir Araujo, Abinew Ali Ayele, Oana Ignat, Alexander Panchenko, and 2 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)*.
- Benjamin Muller, Antonios Anastasopoulos, Benoît Sagot, and Djamé Seddah. 2021. When being unseen from mBERT is just the beginning: Handling new languages with multilingual language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 448–462.
- Neel Nanda, Andrew Lee, and Martin Wattenberg. 2023. Emergent linear representations in world models of self-supervised sequence models. In *Proceedings of the 6th BlackboxNLP Workshop*.
- Emily Öhman, Marc Pàmies, Kaisla Kajava, and Jörg Tiedemann. 2020. XED: A multilingual dataset for sentiment analysis and emotion detection. In *Proceedings of COLING*, pages 6542–6552.
- OLMo Team. 2024. 2 OLMo 2 Furious. Technical report, Allen Institute for AI. <https://huggingface.co/allenai/OLMo-2-1124-7B>.
- Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. 2020. Information-theoretic probing for linguistic structure. In *Proceedings of ACL*, pages 4609–4622.
- Flor Miriam Plaza-del Arco, Alba A. Cercas Curry, Amanda Cercas Curry, and Dirk Hovy. 2024. Emotion analysis in NLP: Trends, gaps and roadmap for future directions. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5696–5710.
- Qwen Team. 2024. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115*.
- Qwen Team. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.
- Lisa Schut, Yarin Gal, and Sebastian Farquhar. 2025. Do multilingual LLMs think in English? *arXiv preprint arXiv:2502.15603*.
- Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Wayne Xin Zhao, Furu Wei, and Ji-Rong Wen. 2024. Language-specific neurons: The key to multilingual capabilities in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5701–5715, Bangkok, Thailand.
- Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of ACL*, pages 4593–4601.
- Andreas Triantafyllopoulos, Johannes Wagner, Hagen Wierstorf, Maximilian Schmitt, Uwe Reichel, Florian Eyben, Felix Burkhardt, and Björn W. Schuller. 2022. Probing speech emotion recognition transformers for linguistic knowledge. In *Proceedings of Interspeech*, pages 146–150.
- Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D’souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya Model: An instruction finetuned open-access multilingual language model. In *Proceedings of ACL*, pages 15894–15939.
- Elena Voita and Ivan Titov. 2020. Information-theoretic probing with minimum description length. In *Proceedings of EMNLP*, pages 183–196.
- Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do Llamas work in English? on the latent language of multilingual transformers. In *Proceedings of ACL*, pages 15366–15394.
- Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of EMNLP-IJCNLP*, pages 833–844.
- Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. 2024. How do large language models handle multilingualism? In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*.