BEYOND LEXICAL ALIGNMENT: TESTING CROSS-LINGUAL COLOUR REPRESENTATIONS AGAINST HUMAN NAMING

Anonymous authors
Paper under double-blind review

ABSTRACT

Languages partition the continuous colour domain into partly different vocabularies, so translation-based alignment alone cannot show whether multilingual representations capture the colours that speakers associate with colour terms. We introduce a human-referenced cross-lingual probe that maps frozen text representations to language-specific human naming centroids for 1,160 colour terms in 14 languages, collected under a common free-naming condition. The primary evaluation holds out each of the ten non-English languages with at least 20 reliable terms from both probe fitting and model selection; a sensitivity analysis including all thirteen non-English languages shows that every model remains above the constant predictor. Five of six models achieve greater relative error reduction than character, token, and frequency controls, and this advantage survives context and exact-string controls in the four models tested. Transfer also persists without English-labelled supervision and across the tested family and script boundaries. The coverage analysis shows where this transfer weakens: standard language holdout leaves the relevant colour regions represented in source supervision, and removing a complete colour region lowers relative error reduction by 0.19–0.25 more than randomly deleting the same number of terms. The results support a transferable mapping from multilingual text representations to human colour referents, while showing that its measured strength depends partly on which regions of the colour domain are represented in source supervision.

1 INTRODUCTION

People speaking different languages often use different inventories of words to divide the same continuous colour domain (Regier et al., 2007; Gibson et al., 2017; Zaslavsky et al., 2018). A multilingual language model faces a nontrivial representational problem: it must distinguish language-specific lexical identities while relating their meanings to colours that can be compared across languages. Translation equivalence alone does not resolve this problem because near translations need not have identical extensions, and many naturally produced colour terms have no one-to-one counterpart.

This tension motivates a more concrete question than whether multilingual embeddings are generally "aligned": can colour information available in some languages predict which colours speakers of a held-out language associate with each word? A positive result would still be ambiguous. A model might place translated strings near one another without matching human referents, or it might encode colour separately in each language, allowing a language-specific decoder to succeed while cross-language transfer fails. Evidence for a shared reference therefore requires both an external human target and a decoder learned without the evaluated language.

Recent work gives increasingly strong evidence that multilingual models share semantic structure. Cross-language token neighbourhoods, representational isomorphism, and linear concept mappings reveal alignment in multilingual embeddings (Wen-Yi & Mimno, 2023; Wu et al., 2024; Peng & Søgaard, 2024); activation analyses and causal patching further support language-agnostic concept representations (Zeng et al., 2025; Dumas et al., 2025). Yet the same literature also identifies important qualifications. Intermediate representations can be biased toward an English latent language (Wendler et al., 2024), transfer can be asymmetric across high- and low-resource languages (Xu et al., 2024), and languages need not lexicalize the same conceptual distinctions (Liu et al., 2023).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Colour offers a useful test case because languages refer to a comparable perceptual domain while dividing it into different lexical categories. Prior work relevant here falls into two strands. One maps English representations to perceptual coordinates or human naming structure (Abdou et al., 2021; Patel & Pavlick, 2022); the other evaluates multilingual alignment through translations or matched lexical concepts. English-only mappings do not test transfer across languages, while translation-based evaluations do not use language-specific human referents as targets. This leaves a question: can a mapping learned from independently sampled vocabularies in other languages predict how speakers of a target language use their own colour words? Shared spelling, an explicitly colour-marked prompt, or English-labelled supervision could otherwise produce apparent transfer without establishing a broadly shared human-referenced signal.

To address this gap, we study frozen text representations of 1,160 naturally produced colour terms. The terms span the 14 languages recorded under the same full-colour free-naming condition. Human responses assign each term a centroid in a common perceptual coordinate system according to the colours for which speakers used it. **The investigation asks three connected questions. (i)** Can a mapping learned from other languages predict a held-out language's human naming centroids beyond form and frequency baselines? **(ii)** If so, does that transfer require English-labelled or same-family examples? **(iii)** How much transfer remains when source supervision excludes both the target language and the colour region containing its referents? The questions progress from establishing transfer to identifying its source-language requirements and then to testing whether ordinary language holdout truly evaluates generalization beyond regions already represented in source supervision. Context, exact-string, and random-correspondence controls address direct lexical alternatives to the transfer result.

The findings narrow the interpretation in three steps. First, five of six models predict held-out-language referents beyond surface baselines, and the advantage survives the tested context and exact-string controls. Second, the transfer persists without English-labelled supervision and across the tested language-family and script boundaries. Third, ordinary language holdout still leaves nearby colours represented by source-language terms. Removing a complete target region causes a larger loss than equal-count random deletion, although larger models remain predictive. The evidence supports a transferable mapping to human colour referents, with measured performance partly depending on referential coverage.

Contributions. We contribute (i) a controlled evaluation that grounds cross-language representation analysis in language-specific human observations; (ii) evidence that a representation-to-reference mapping transfers without English-labelled or same-family supervision in the tested settings; and (iii) a referential-region holdout that separates transfer across vocabularies from generalization beyond covered parts of the target domain. These contributions connect multilingual lexical alignment to independently measured human reference and show that language holdout establishes lexical transfer, but not necessarily prediction in unseen colour regions.

2 Related Work

Human colour naming as an external reference. Human colour naming offers a common perceptual reference without assuming that languages use the same lexical categories. Classical work explains recurring patterns and variation in colour vocabularies through perceptual constraints, communicative need, and efficient partitioning (Regier et al., 2007; Gibson et al., 2017; Zaslavsky et al., 2018). This cross-linguistic variation makes translation equivalence an incomplete criterion for semantic correspondence: languages can apply approximate equivalents to different portions of the colour domain. Computational studies have related English colour-word representations to human naming and perceptual coordinates through representational similarity and learned mappings (Abdou et al., 2021), and have mapped language-model representations to grounded conceptual spaces (Patel & Pavlick, 2022). These studies establish that a representation-to-colour relation can be recovered within a source language. They do not establish that the same relation generalizes across independently elicited vocabularies. We extend this line of work by estimating the reference separately from each language's naming responses and making cross-language transfer of the mapping the primary test.

Cross-lingual alignment, shared concepts, and language asymmetry. Multilingual representation research finds cross-language structure in token neighbourhoods (Wen-Yi & Mimno, 2023),

approximately isomorphic sentence spaces (Wu et al., 2024), and linear mappings between corresponding lexical concepts (Peng & Søgaard, 2024). Activation-level studies likewise report increasingly language-agnostic semantic patterns (Zeng et al., 2025), while causal patching can manipulate concept and output language separately in translation (Dumas et al., 2025). These results provide strong evidence for shared structure, but semantic correspondence in these studies is usually defined through translations, parallel meanings, or task labels. Because languages need not lexicalize identical conceptual distinctions (Liu et al., 2023), alignment between matched words does not by itself show that model representations preserve how speakers of each language use those words. Our human centroids change the criterion from agreement between lexical counterparts to prediction of independently observed referents.

Shared structure may also privilege particular languages. Middle-layer predictions can favour English before returning to the input language (Wendler et al., 2024), and transfer can be asymmetric across languages with different resource levels (Xu et al., 2024). These findings directly motivate our source-language diagnostics. Removing English-labelled examples tests whether English supervision is necessary for fitting the readout, while Indo-European-to-Korean/Chinese transfer tests whether same-family supervision is necessary in the examined cases. These tests concern the supervision used to fit the readout. They do not identify the model's latent routing mechanism.

What cross-language probing establishes. A successful probe shows that information is recoverable by a specified function class; it does not by itself establish how that information is geometrically organized or causally used (Hewitt & Liang, 2019). The inference is weaker still if form or data-distribution cues can predict the target. For colour words, character similarity, shared tokenizer units, frequency, identical spellings, and an explicitly colour-marked context are plausible alternatives. Matched feature baselines, context comparison, and exact-string removal address these alternatives. A separate random-correspondence check tests whether the fixed predictions remain associated with the correct human targets. The cross-language split then asks whether the representation-to-reference relation is recoverable without fitting or selecting on the evaluated language.

Language holdout also leaves another source of information uncontrolled: other languages may label the same part of the reference space. Spatial block holdout has been used to distinguish interpolation from extrapolation in learned representations (Gurnee & Tegmark, 2024). We adapt this logic to the human colour domain by comparing removal of a k-means referential region with removal of the same number of terms at random. This adds a split criterion absent from ordinary cross-lingual alignment tests: whether the target's referential neighbourhood, rather than only its language, was observed during readout fitting.

This representation-level question is distinct from recent behavioural evaluations. ColorBench supplies visual inputs and tests perception, reasoning, and robustness in vision–language models (Liang et al., 2025); sensory benchmarks compare model judgments with human data across modalities (Marjieh et al., 2024); and multilingual colour studies evaluate generated names or translations (Feldman & Peng, 2026; Mylonas et al., 2026). In particular, broad lexical availability can coexist with poor agreement near low-chroma colours and human category boundaries (Feldman & Peng, 2026). Our text-only models receive colour words, and an external readout predicts their typical human referents from frozen activations. Our task sits between these two evaluation levels. It tests a cross-language mapping from text representations to human colour referents, rather than generated names or translation behaviour.

3 METHODOLOGY

3.1 OVERVIEW

We use a linear readout trained without target-language data to test whether frozen residual-stream representations make human colour referents recoverable across languages. Human naming responses first locate each colour term in a common perceptual space. A frozen language model then supplies a text representation of that term, and a readout learned only from other languages predicts its human colour target. Figure 1 shows this sequence and the three comparisons used to interpret successful prediction.

Language holdout establishes transfer across lexical systems, but it does not reveal what makes that transfer possible. The remaining languages may support a shared representation-to-reference

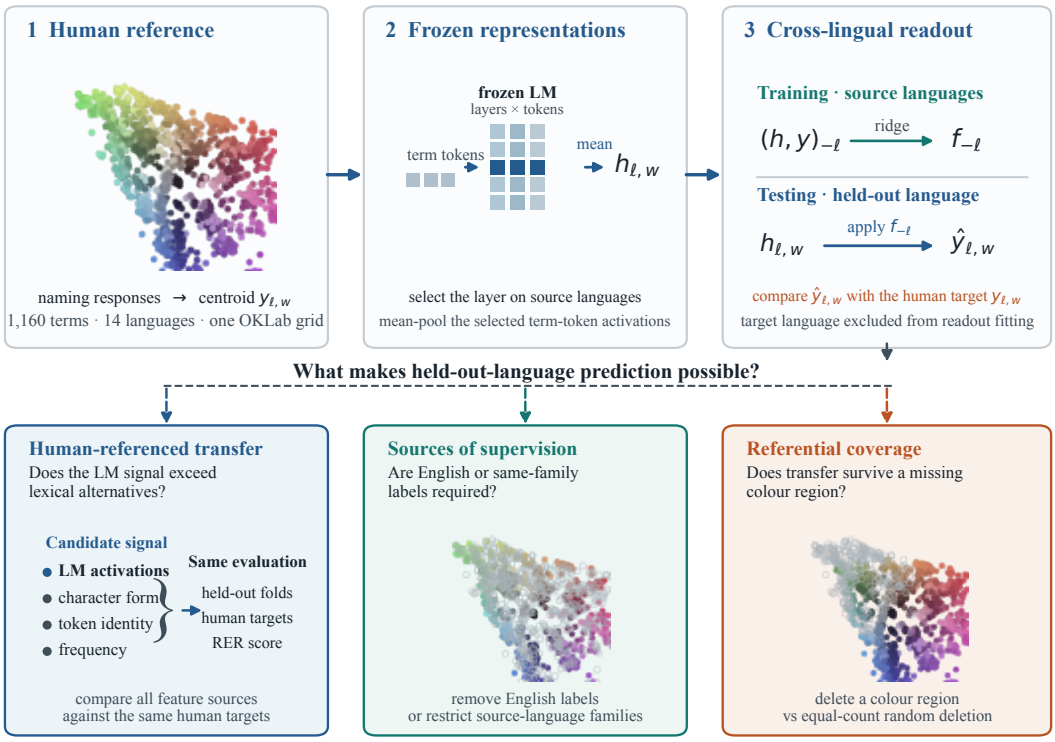

Figure 1: Evaluation framework. Human naming responses define language-specific targets in a common colour space; a frozen language model supplies text representations; and a linear readout fitted on source languages predicts a held-out language. The analyses then test lexical alternatives, dependence on particular supervising languages, and dependence on referential coverage. Hollow points denote supervision removed in a diagnostic condition.

relation. Alternatively, they may cover the same colour regions densely enough to support prediction near familiar referents. The evaluation moves from the primary transfer test to restrictions on the source languages and finally to removal of an entire region of the human reference space.

3.2 Human reference targets in OKLab

Each naming observation pairs a term w in language ℓ with the colour stimulus for which a speaker produced it. We express these stimuli in OKLab, a three-dimensional space designed to make coordinate distance more consistent with perceptual difference than distance in RGB (Ottosson, 2020). Its axes encode lightness (L), green–red position (a), and blue–yellow position (b).

Let $p_{\ell,w}(c)$ be the normalized human naming distribution for term w over colour points c on the common OKLab grid. The target is its expected coordinate,

$$y_{\ell,w} = \sum_{c} p_{\ell,w}(c)c \in \mathbb{R}^3. \tag{1}$$

This centroid represents the term's typical referent in the observed naming data. It permits comparison across languages without assuming that translations have identical meanings. Terms receive reliability weights from participant split-half stability; Section 4.1 gives the construction and inclusion rules.

3.3 Frozen representations and cross-lingual linear readout

For the same term, the frozen language model supplies a residual-stream representation $h_{\ell,w} \in \mathbb{R}^d$. When language ℓ is evaluated, none of its terms or human targets is used to fit the readout or choose its hyperparameters. We learn $f_{-\ell} : \mathbb{R}^d \to \mathbb{R}^3$ from the remaining languages and compare $f_{-\ell}(h_{\ell,w})$ with $y_{\ell,w}$.

The readout is weighted ridge regression,

$$\widehat{W} = \arg\min_{W} \sum_{i} w_i \|Wh_i - y_i\|_2^2 + \alpha \|W\|_F^2. \tag{2}$$

Representations are centred and scaled using source-language data, while targets are centred using the source-language weighted mean. Each included language receives equal total training weight. An inner leave-one-language-out loop first selects the ridge penalty and then the representation layer. The readout is then refitted on all permissible source languages. Because the language model is frozen and the mapper is affine, successful transfer indicates a shared linear relation available in its representations rather than colour knowledge learned by the probe from the target language.

3.4 Relative error reduction

Euclidean distance is the straight-line distance in OKLab between a predicted centroid and its human target; smaller values are better. We compare the readout's weighted prediction error, E_{readout}, with the error of a simple constant colour predictor, E_{constant}. The constant predictor assigns every test term the same colour: the weighted mean centroid of the source-language training terms. We call the resulting quantity the relative error-reduction (RER) score:

$$\text{RER} = 1 - \frac{E_{\text{readout}}}{E_{\text{constant}}} = 1 - \frac{\sum_i w_i \|\widehat{y}_i - y_i\|_2}{\sum_i w_i \|\overline{y}_{\text{train}} - y_i\|_2}. \tag{3}$$

A score of 0 means that the readout makes the same error as the constant predictor; 0.5 means that it halves that error; 1 denotes perfect centroid prediction; and a negative value is worse than the constant predictor. This measure is not R^2. The primary score averages the ten evaluated languages equally.

4 Experiments

4.1 Experimental setup

Data, targets, and language splits. We use a fixed snapshot of the UW *Many Languages, Many Colors* multilingual colour-naming dataset (Kim et al., 2019; UW Interactive Data Lab, 2026). The repository extends the 2019 publication data with later observations. We retain the full-colour RGB condition on a white background for all 14 available languages. In the free-naming task, participants viewed colour tiles and supplied unconstrained names in their native language. We chose this dataset for three reasons: every response pairs a naturally produced term with a known colour stimulus; the languages share the same task and stimulus domain; and their vocabularies were elicited independently rather than constructed through translation. This design provides language-specific human referents that can be compared in a common colour space without assuming one-to-one lexical equivalence.

We convert the stimuli to a shared grid of 2,048 in-gamut OKLab points. Gaussian smoothing with bandwidth 0.06 estimates each term's naming distribution, with kernels truncated at three bandwidths. We then divide by the smoothed count of all responses in the language to reduce dependence on where colours were sampled. The normalized distribution yields the centroid defined in Section 3.2.

The validated lexicon contains 1,160 terms in 14 languages: 595 English terms, 139 Korean, 84 Chinese, and between 5 and 56 in each other language. A term requires at least 12 naming records and at least eight estimable results across ten participant split-halves. Reliability is inversely related to median split-half centroid distance and is clipped at the within-language fifth and ninety-fifth percentiles. Appendix C reports target stability.

All 14 languages can provide source terms. The primary evaluation includes the ten non-English languages with at least 20 retained terms, a threshold fixed before model extraction. Romanian (5), Finnish (13), and Swedish (19) remain in the training pool but are excluded from the primary language macro; English is excluded because the target claim concerns transfer to non-English languages and this permits a common English-removal analysis. Appendix B lists all languages, and Appendix C reports sensitivity over all thirteen non-English languages.

Models and representations. We evaluate Qwen2.5-0.5B, 1.5B, and 7B; Llama-3.1-8B; Gemma-2-9B; and Qwen3-14B-Base. All models remain frozen. The primary input is `Color: <term>` in every language. We average residual-stream activations over the term's tokens, excluding the carrier and special tokens, and store them in float16. Human colour coordinates are never supplied to the language model. Appendix H reports token-span audits.

Probe selection, baselines, and controls. For each outer fold, we first remove the target language. Inner language holdout selects $\alpha \in \{0.01, 0.1, 1, 10, 100\}$ at the final layer and then selects the layer with that penalty fixed. We finally refit the readout on all source languages. Character n-grams, tokenizer identities, corpus frequency, and characters plus frequency use the same cross-language ridge evaluation. Replacing `Color:` with `Word:`, removing exact string matches, and permuting target correspondence test specific lexical alternatives. Appendix D gives these controls. Appendix E reports sensitivity to language composition, and Appendix I compares raw cosine geometry with human colour distances.

Source and coverage diagnostics. The source-language diagnostics compare full supervision with four alternatives: removing English, reducing English to 35 terms, removing 595 terms proportionally across languages, and using English alone. We also evaluate Korean and Chinese using only Indo-European supervision. The coverage diagnostic partitions the human centroids into six k-means regions (ten initializations; seed 0). For each evaluated region, we remove every source term in that region and compare the result with ten equal-count random removals. Only language–region cells with at least five test terms contribute. The six-model comparison retains the layer and penalty selected under the full protocol. A two-model sensitivity check repeats both selections after region removal. The partition is target-informed because it uses human centroids, including evaluation targets, but it does not use model activations.

We use a nearest-neighbour measure to quantify how much of the target region is already represented before deletion. Coverage is the fraction of test centroids whose nearest permissible source centroid lies within OKLab threshold t. Appendix G reports $t \in \{0.02, 0.04, 0.06\}$ and the region-level analyses.

4.2 Cross-lingual referent transfer

The first experiment asks whether a readout learned entirely from other languages predicts the held-out language's human colour referents beyond surface-form alternatives.

Table 1: Primary macro relative error reduction (RER) for human colour-centroid prediction in ten non-English held-out languages. Values are point estimates averaged over languages; 0 matches the constant source-mean predictor and 1 denotes perfect prediction. Each model uses the same nested evaluation protocol; language-composition sensitivity intervals are reported in Appendix E.

Surface-feature baselines						
Representation	Char. n-grams	Subwords	Log freq.	Char. + freq.		
Error reduction	.186	.080	.003	.191		
Language-model representations						
	Qwen2.5	Qwen2.5	Qwen2.5	Llama-3.1	Gemma-2	Qwen3-14B
Representation	0.5B	1.5B	7B	8B	9B	Base
Error reduction	.169	.372	.474	.520	.545	.571

A linear mapping learned from other languages predicts held-out-language human colour centroids (Table 1). Five models exceed the strongest surface baseline (0.191), reducing error by 0.372–0.571 relative to the constant predictor. Qwen2.5-0.5B improves on the constant predictor but not on characters plus frequency. The result also survives the tested lexical controls. Changing the carrier from `Color:` to `Word:` preserves positive prediction in all four tested models, and exact-string exclusion preserves their advantage over matched character features. Appendix D reports these controls and the random-correspondence check. Appendix F reports per-language variation.

These results establish human-referenced transfer under language holdout. The next experiment asks whether that transfer depends on a particular supervising language or a related language family.

4.3 Transfer without English or same-family supervision

English supplies 51.3% of the lexicon, making it the strongest candidate source of apparent transfer from a single language. Because languages receive balanced training weight, the size of the English lexicon does not directly determine its contribution to fitting. Removing English-labelled examples tests whether the readout nevertheless depends on them. Table 2 reports this comparison and the matched source-pool controls.

Table 2: Changing the source of supervision. All entries are primary language-macro relative error reduction. The first three columns are Qwen2.5 models. Random deletion controls term count, but not colour coverage or language composition.

Condition	Qwen2.5 0.5B	Qwen2.5 1.5B	Qwen2.5 7B	Llama-3.1 8B	Gemma-2 9B	Qwen3-14B Base
Full	.169	.372	.474	.520	.545	.571
Without English	.133	.354	.471	.521	.552	.548
English reduced to 35	.139	.356	.469	.517	.549	.554
Size-matched removal	.113	.318	.453	.482	.538	.529
English only	.138	.279	.404	.427	.509	.429

Removing English produces smaller losses than equal-count proportional deletion in all six models (Table 2). All models remain predictive, and the four largest stay close to full supervision. Useful prediction remains without English-labelled examples.

Restricting supervision to 924 Indo-European terms leaves Qwen2.5-7B nearly unchanged on Korean and Chinese and improves Llama-3.1-8B on both (Figure 2). The readout transfers across both language-family and writing-system differences in these tested cases.

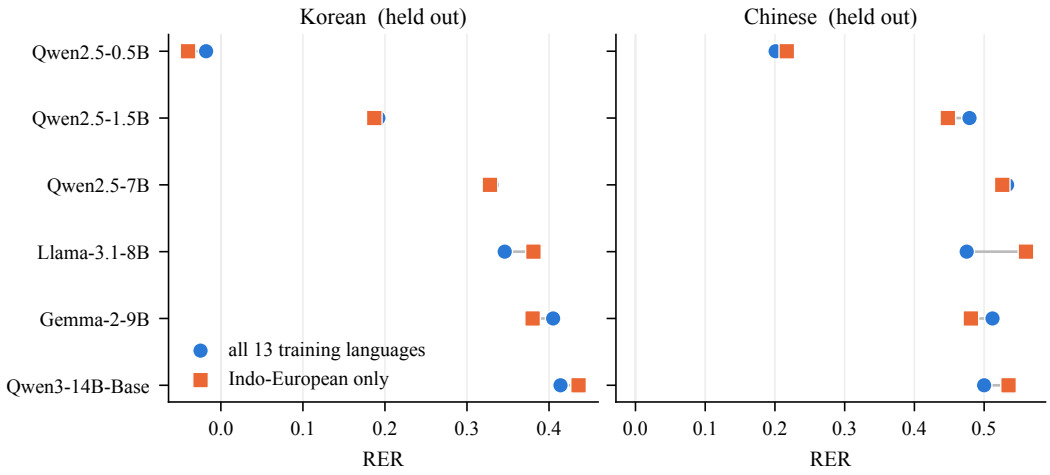

Figure 2: Full-pool and Indo-European-only supervision for Korean and Chinese. Points are RER estimates with no uncertainty intervals plotted. Similar restricted-pool scores show that same-family labels are unnecessary in these tested cases.

These tests show that the observed transfer does not require an English-labelled bridge or same-family supervision in the examined cases. They do not show whether the remaining languages supply nearby referential anchors, which motivates the final analysis.

4.4 Referential coverage limits region-level generalization

Holding out a language does not necessarily hold out the colours its terms denote. At the 0.04 threshold, every test term has a source-language neighbour in OKLab. Ordinary language holdout evaluates new words while leaving their colour regions covered by other languages. We compare deletion of an entire target region with random deletion of the same number of source terms. If only training-set size matters, the two conditions should have similar effects. If local referential coverage matters, region deletion should be more damaging.

Table 3: Relative error reduction within a held-out target region. Region-macro scores average eligible languages within each k-means region and then weight the six regions equally; they are not directly comparable to the language-macro scores in Table 1. Region deletion is consistently worse than equal-count random deletion; the final column gives their difference.

Model	Full	Random removal	Region removal	Difference
Qwen2.5-0.5B	.104	.100	−.152	−.252
Qwen2.5-1.5B	.274	.265	.028	−.237
Qwen2.5-7B	.378	.374	.157	−.217
Llama-3.1-8B	.394	.391	.194	−.196
Gemma-2-9B	.449	.452	.261	−.191
Qwen3-14B-Base	.493	.484	.253	−.232

Under the fixed full-protocol settings, region removal lowers relative error reduction by 0.19–0.25 compared with random removal. By contrast, random deletion changes the full-pool score by less than 0.01 in every model (Table 3). Re-selecting the layer and penalty after removal preserves a roughly 0.19 penalty in both Qwen2.5-7B and Llama-3.1-8B, with paired intervals below zero (Table 17). This contrast shows that the location of supervision in the colour domain matters beyond the number of available terms. Larger models remain better than the constant predictor after removal, so referential coverage supports transfer but does not fully account for it. A complementary raw-distance diagnostic finds weaker cross-language correspondence at the selected layers: 0.000–0.115 without weighting and 0.019–0.198 with reliability weighting. Appendix I reports these coefficients on their own scale. Appendix G reports the language-level coverage profile, region variation, and training-size sensitivity.

5 Discussion

Our results distinguish two forms of generalization that standard cross-lingual evaluations often conflate: transfer to a new language and prediction in a colour region absent from supervision. Multilingual representations support a shared linear mapping to human colour referents, but the strength of that mapping depends partly on how densely the probe's source languages cover the target referential region.

5.1 Human reference changes what counts as cross-lingual alignment

Translation-based alignment defines correspondence through matched words or concepts. Human naming supplies a different criterion: each term is evaluated against the colours for which speakers of that language actually produced it. A single affine readout learned without the target language predicts these independently measured referents. This extends English-only mappings between language-model representations and colour structure (Abdou et al., 2021; Patel & Pavlick, 2022). It also changes the multilingual criterion from agreement between lexical counterparts to prediction of an external referential variable (Wu et al., 2024; Peng & Søgaard, 2024).

The contrast with raw geometry clarifies what this result establishes. At the selected layers, cosine distances correspond only modestly with distances between human naming centroids, whereas the shared linear readout achieves substantially higher RER scores than form and frequency baselines. The result suggests that colour reference is recoverable from a shared cross-language subspace, even though it is not dominant in the overall cosine geometry. This is evidence for recoverable referential structure, not for a complete human-like colour space. The centroids likewise capture typical referents rather than full category boundaries, complementing evidence that broad lexical availability can coexist with disagreement near low-chroma colours and human category boundaries (Feldman & Peng, 2026).

This interpretation applies to centroid prediction rather than to full naming distributions. An exploratory distributional diagnostic in Appendix K does not support extending the claim to recovery of those distributions. Because that analysis uses human-derived nuisance features, we report it separately from the primary evaluation.

From a cognitive-linguistic perspective, this pattern shows how shared constraints and language-specific categorization can coexist in a model representation. Languages divide the same perceptual domain into partly different lexical categories (Regier et al., 2007; Gibson et al., 2017; Zaslavsky

et al., 2018); the readout transfers across those partitions without requiring approximate translations to share an identical target. Multilingual representations may connect complementary ways of naming a common domain without erasing their language-specific referential differences.

5.2 Language transfer is not referential extrapolation

Removing English-labelled examples leaves substantial prediction, and Indo-European supervision predicts Korean and Chinese. The external mapping can be learned from different linguistic samples of the colour domain rather than requiring English or same-family labels in the examined cases. This result complements evidence for typological advantages, English-pivot dynamics, and resource-linked asymmetries (Peng & Søgaard, 2024; Wendler et al., 2024; Xu et al., 2024). Model computation may remain asymmetric even when a particular language is unnecessary for supervising the readout. Our experiment does not distinguish a single language-neutral interlingua from aligned language-specific spaces. It shows more narrowly that successful external transfer does not require one labelled language to serve as the canonical bridge.

Language holdout, however, leaves nearby human referents densely represented by other languages. The high score may reflect both transfer across lexical systems and interpolation among colour regions already represented by source-language terms. Region removal causes a much larger loss than equal-count random deletion, showing that the location of supervision matters beyond its quantity. Yet larger models remain predictive after region removal, so local coverage is an enabling condition rather than a complete explanation. Lexical transfer and referential extrapolation are distinct generalization problems.

5.3 Referential coverage as an evaluation principle

Cross-lingual evaluations should describe not only which languages are held out, but also which parts of the target domain remain covered by the supervising languages. When an external reference space is available, a language split can test transfer across lexical systems and a region split can separately test generalization beyond covered referents. Reporting both prevents dense semantic overlap from being mistaken for unconstrained cross-domain generalization, a distinction that translation-based benchmarks usually leave implicit.

The same distinction suggests two extensions. For low-resource alignment, coverage-aware source selection may be more useful than adding further examples from referential regions that are already well represented. For multimodal models, the same region-holdout design could test whether visual grounding reduces dependence on textual coverage and improves prediction in uncovered parts of the colour domain (Liang et al., 2025; Marjieh et al., 2024). These are hypotheses for future evaluation; the present experiments concern frozen text representations rather than controlled multilingual training or visual grounding.

6 Conclusion

Colour naming offers a shared perceptual domain in which lexical meanings can be compared without assuming that languages divide the domain in the same way. Using human naming centroids from the UW multilingual colour-naming data, we tested whether a linear readout learned from non-target languages could predict the referents of colour terms in a held-out language. Five of six models outperformed character, token, and frequency baselines. The advantage survived the tested context and exact-string controls and persisted without English-labelled supervision and across the tested family and script boundaries. Yet standard language holdout leaves the relevant colour regions represented by terms in the source languages: removing a complete colour region from source supervision caused a larger loss than deleting the same number of terms at random. The results support a transferable mapping from multilingual text representations to human colour referents, while showing that its measured strength depends partly on which regions of the colour domain are represented in source supervision. Language holdout tests transfer across lexical systems, but it does not by itself test whether the model can predict colours in regions absent from source supervision.

Limitations. Our analysis is restricted to text-only representations and human naming centroids. It evaluates cross-language reference mapping under the tested language, model, and coverage conditions, rather than visual colour perception or a complete human-like colour space.

Reproducibility Statement

Appendix B documents the lexicon and selected probe settings; subsequent appendices report target stability, controls, uncertainty, per-language results, and coverage analyses. The accompanying materials retain the target-construction, extraction, nested-selection, and diagnostic scripts needed to reproduce the reported comparisons. An anonymous supplementary archive provides the core code and compact result files for these analyses.

AI Use Statement

Generative AI tools were used for language editing, manuscript organization, and figure-layout iteration. All experimental design, code implementation, execution, and reported results were independently carried out and verified by the author.

References

Mostafa Abdou, Artur Kulmizev, Daniel Hershcovich, Stella Frank, Ellie Pavlick, and Anders Søgaard. Can language models encode perceptual structure without grounding? a case study in color. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pp. 109–132, 2021. doi: 10.18653/v1/2021.conll-1.9.

Clément Dumas, Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. Separating tongue from thought: Activation patching reveals language-agnostic concept representations in transformers. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 31822–31841, 2025. doi: 10.18653/v1/2025.acl-long.1536.

Anna Feldman and Jing Peng. Lexical availability and human distributional agreement in GPT-4o's color naming. In *Proceedings of the 15th Joint Conference on Lexical and Computational Semantics (*SEM 2026)*, pp. 136–147, 2026. doi: 10.18653/v1/2026.starsem-conference.9.

Edward Gibson, Richard Futrell, Julian Jara-Ettinger, Kyle Mahowald, Leon Bergen, Sivalogeswaran Ratnasingam, Mitchell Gibson, Steven T. Piantadosi, and Bevil R. Conway. Color naming across languages reflects color use. *Proceedings of the National Academy of Sciences*, 114(40):10785–10790, 2017. doi: 10.1073/pnas.1619666114.

Wes Gurnee and Max Tegmark. Language models represent space and time. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.02207.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP)*, 2019. doi: 10.18653/v1/D19-1275.

Younghoon Kim, Kyle Thayer, Gabriella Silva Gorsky, and Jeffrey Heer. Color names across languages: Salient colors and term translation in multilingual color naming models. In *EuroVis 2019 – Short Papers*, 2019. doi: 10.2312/evs.20191166.

Yijun Liang, Ming Li, Chenrui Fan, Ziyue Li, Dang Nguyen, Kwesi Cobbina, Shweta Bhardwaj, Jiuhai Chen, Fuxiao Liu, and Tianyi Zhou. Colorbench: Can vlms see and understand the colorful world? a comprehensive benchmark for color perception, reasoning, and robustness. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.nips.cc/paper_files/paper/2025/hash/d34497330b1fd6530f7afd86d0df9f76-Abstract-Datasets_and_Benchmarks_Track.html.

Yihong Liu, Haotian Ye, Leonie Weissweiler, Philipp Wicke, Renhao Pei, Robert Zangenfeind, and Hinrich Sch"utze. A crosslingual investigation of conceptualization in 1335 languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12969–13000, 2023. doi: 10.18653/v1/2023.acl-long.726.

Raja Marjieh, Ilia Sucholutsky, Pol van Rijn, Nori Jacoby, and Thomas L. Griffiths. Large language models predict human sensory judgments across six modalities. *Scientific Reports*, 14:21445, 2024. doi: 10.1038/s41598-024-72071-1.

Dimitris Mylonas, Rafique Ahmed, Akvile Sinkeviciute, and Alexandros Koliousis. Colour in translation: Data, models, and benchmarking for cross-linguistic colour naming. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, 2026. doi: 10.1145/3772318.3791626.

Björn Ottosson. A perceptual color space for image processing, 2020. URL https://bottosson.github.io/posts/oklab/.

Roma Patel and Ellie Pavlick. Mapping language models to grounded conceptual spaces. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=gJcEM8sxHK.

Qiwei Peng and Anders Søgaard. Concept space alignment in multilingual LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5511–5526, 2024. doi: 10.18653/v1/2024.emnlp-main.315.

Terry Regier, Paul Kay, and Naveen Khetarpal. Color naming reflects optimal partitions of color space. *Proceedings of the National Academy of Sciences*, 104(4):1436–1441, 2007. doi: 10.1073/pnas.0610341104.

UW Interactive Data Lab. The many languages, many colors project, 2026. URL https://idl.uw.edu/color-naming-in-different-languages/. Dataset and project website, updated September 5, 2026.

Andrea W. Wen-Yi and David Mimno. Hyperpolyglot LLMs: Cross-lingual interpretability in token embeddings. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 1124–1131, 2023. doi: 10.18653/v1/2023.emnlp-main.71.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. Do llamas work in English? on the latent language of multilingual transformers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024. doi: 10.18653/v1/2024.acl-long.820.

Di Wu, Yibin Lei, Andrew Yates, and Christof Monz. Representational isomorphism and alignment of multilingual large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024. URL https://aclanthology.org/2024.findings-emnlp.823/.

Shaoyang Xu, Weilong Dong, Zishan Guo, Xinwei Wu, and Deyi Xiong. Exploring multilingual concepts of human values in large language models: Is value alignment consistent, transferable and controllable across languages? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 1771–1793, 2024. doi: 10.18653/v1/2024.findings-emnlp.96.

Noga Zaslavsky, Charles Kemp, Terry Regier, and Naftali Tishby. Efficient compression in color naming and its evolution. *Proceedings of the National Academy of Sciences*, 115(31):7937–7942, 2018. doi: 10.1073/pnas.1800521115.

Hongchuan Zeng, Senyu Han, Lu Chen, and Kai Yu. Converging to a lingua franca: Evolution of linguistic regions and semantics alignment in multilingual large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 10602–10617, 2025. URL https://aclanthology.org/2025.coling-main.707/.

A Reproducibility details

The study uses public human naming records and frozen model activations. The project retains target construction, lexicon validation, extraction, nested selection, supervision diagnostics, per-language results, and input hashes. Full-condition outputs are checked against the original selected probes; region, target-stability, distance, and uncertainty summaries are regenerated from saved results or activations.

B Lexicon and selected probe settings

The table lists all training-pool languages. The primary evaluation macro excludes English and the three marked languages. The selection table gives representative model families; both layer and regularization are selected without the outer test language. The regularization grid and sequential selection procedure are specified in Section 3.

Table 4: The frozen lexicon. *terms* is the count entering all analyses; *records* is the median number of human naming responses behind a term's centroid. Languages below the $n \geq 20$ threshold are marked *.

code	language	terms	median records	family	script
en	English	595	34	Indo-European	Latn
ko	Korean	139	28	other	Hang
zh	Chinese	84	30	other	Hani
de	German	56	30	Indo-European	Latn
es	Spanish	51	27	Indo-European	Latn
fa	Persian	45	28	Indo-European	Arab
fr	French	45	31	Indo-European	Latn
ru	Russian	35	31	Indo-European	Cyrl
pt	Portuguese	28	32	Indo-European	Latn
pl	Polish	23	25	Indo-European	Latn
nl	Dutch	22	32	Indo-European	Latn
sv*	Swedish	19	22	Indo-European	Latn
fi*	Finnish	13	21	other	Latn
ro*	Romanian	5	22	Indo-European	Latn
total	14 languages	1160			

Table 5: Layer and regularisation selected by the inner leave-one-language-out loop, per outer fold. Selection never sees the held-out language.

	Qwen2.5-7B		Llama-3.1-8B		Qwen3-14B-Base	
held out	layer	α	layer	α	layer	α
de	15	1	16	0.1	12	0.1
en	9	1	10	0.1	12	0.01
es	9	1	8	0.1	13	0.01
fa	8	0.1	16	0.1	13	0.1
fi	7	0.1	10	0.1	13	0.01
fr	7	0.1	16	0.1	13	0.01
ko	11	0.1	8	0.1	12	0.1
nl	7	0.1	16	0.1	12	0.01
pl	8	0.1	16	0.1	13	0.01
pt	9	1	16	0.1	13	0.01
ro	10	0.1	16	0.1	13	0.01
ru	8	0.1	11	0.1	13	0.01
sv	15	0.1	16	0.1	13	0.01
zh	9	1	7	0.1	12	0.1

C HUMAN TARGET QUALITY AND SENSITIVITY

Table 6: Human target stability. For each term, we first take its median centroid distance across valid participant splits; quartiles below summarize those term-level medians. Distances are in OKLab.

Language	Terms	Median records	Distance Q25	Median	Q75
de	56	30	0.022	0.036	0.056
en	595	34	0.017	0.029	0.044
es	51	27	0.027	0.043	0.068
fa	45	28	0.023	0.031	0.042
fi	13	21	0.033	0.045	0.054
fr	45	31	0.031	0.051	0.074
ko	139	28	0.017	0.030	0.044
nl	22	32	0.022	0.034	0.048
pl	23	25	0.033	0.039	0.052
pt	28	32	0.024	0.029	0.044
ro	5	22	0.070	0.074	0.104
ru	35	31	0.023	0.032	0.051
sv	19	22	0.027	0.035	0.053
zh	84	30	0.022	0.035	0.054

Table 7: Sensitivity of the model RER score to the language-level inclusion threshold. The primary macro averages ten non-English languages with at least 20 retained terms. The second column additionally includes Romanian (5 terms), Finnish (13), and Swedish (19). The headline model–baseline comparison uses the primary ten-language population.

Model	Primary 10	All 13 non-English
Qwen2.5-0.5B	.169	.170
Qwen2.5-1.5B	.372	.353
Qwen2.5-7B	.474	.418
Llama-3.1-8B	.520	.469
Gemma-2-9B	.545	.514
Qwen3-14B-Base	.571	.536

The quartiles describe variation across retained words, not confidence intervals. Participant splitting excludes records without known participant identifiers from the split comparison. Reliability weighting emphasizes more stable targets but cannot remove systematic sampling bias.

Existing sensitivity runs restrict both training and test lexicons to terms with at least 20 naming records (828 terms overall) while retaining the original fold-specific layer and regularization. Averaging the same ten primary languages yields RER scores of 0.163, 0.371, 0.483, and 0.519 for Qwen2.5-0.5B, Qwen2.5-1.5B, Qwen2.5-7B, and Llama-3.1-8B, respectively; some languages have fewer than 20 remaining test terms. The corresponding main scores are 0.169, 0.372, 0.474, and 0.520. Predictive performance persists on the higher-record subset. Because the training and test populations and constant predictors change, this analysis does not isolate target noise or establish robustness of the English-removal contrast. Matching restricted-set feature baselines and English-removal runs are not available.

D CONTEXT AND EXACT-STRING CONTROLS

The context comparison uses exactly the primary evaluation terms, reliability weights, and constant denominators. The saved neutral-context runs refit the probe using the layer and regularization selected under `Color:`. They do not reselect those settings for `Word:`. The contrast measures sensitivity under the original settings rather than the best possible performance for either prompt. Both carriers are English, so the result does not rule out English contextual influence.

For exact-string exclusion, training and model selection remain unchanged. A target term is excluded if its exact stored string occurs in any other language's training lexicon. The same 473 retained words are used for the probe and character baseline, with weights renormalized within each

Table 8: Existing controls summarized over the ten primary languages. Context columns use all evaluation terms; exact-string exclusion columns use the same restricted test words for both methods and a common full-training-mean constant. Models are refitted for the neutral context at the original selected settings; string exclusion only re-evaluates saved predictions.

	Context		Exact-string exclusion	
Model	`Color:`	`Word:`	Probe	Characters
Qwen2.5-0.5B	0.169	0.164	0.141	0.115
Qwen2.5-1.5B	0.372	0.337	0.353	0.115
Qwen2.5-7B	0.474	0.449	0.446	0.115
Llama-3.1-8B	0.520	0.454	0.492	0.115

language. Saved character errors are divided by the same restricted-test error of the full-training-mean constant as the model errors. The evaluation population differs from the full test set, so changes from the full-set RER score cannot be attributed solely to overlap. This check excludes identical strings, not differently spelled cognates, borrowings, or semantic associations. Restricted-set character-plus-frequency results are not supplied; the comparison here is specifically with characters.

Fixed-prediction permutation check. The existing check permutes saved held-out predictions against fixed target–weight rows within each language, using 10,000 permutations. It averages Euclidean error equally over all fourteen languages, including English and the small lexicons, and is distinct from the primary ten-language RER aggregate. All four evaluated models, including Qwen2.5-0.5B, attain $p = 1/10{,}001$ for lower error than random correspondence. This tests the association of fixed predictions with the correct targets. It neither repeats layer selection or fitting under a shuffled-training null nor demonstrates superiority to the lexical baselines. In particular, the smallest model's positive association does not make it a known no-signal control.

E Sensitivity of main comparisons to language composition

For each comparison, we retain its paired scores in the ten primary languages, sample ten language indices with replacement, and average the paired differences. We use 20,000 draws (seed 20260919) and report the 2.5th and 97.5th percentiles. Pairing preserves each language's contribution to both conditions. Models, targets, predictions, and within-language weights remain fixed. These descriptive intervals quantify sensitivity to reweighting the observed languages. They are not calibrated confidence intervals for arbitrary languages because the languages were not randomly sampled and the outer folds share training data. They also exclude participant-sampling and probe-refitting uncertainty. We do not infer significance or equivalence from whether an interval includes zero.

Table 9: Paired language-composition sensitivity: point differences and central 95% resampling intervals. These are not full-pipeline confidence intervals or equivalence tests. Positive values favour the model over the surface baseline; negative English-removal changes indicate a loss.

Model	Model minus characters+frequency	Without English minus full
Qwen2.5-0.5B	-0.022 [-0.099, +0.072]	-0.036 [-0.061, -0.010]
Qwen2.5-1.5B	+0.181 [+0.093, +0.288]	-0.018 [-0.033, -0.003]
Qwen2.5-7B	+0.283 [+0.201, +0.383]	-0.003 [-0.018, +0.014]
Llama-3.1-8B	+0.329 [+0.265, +0.400]	+0.000 [-0.024, +0.024]
Gemma-2-9B	+0.354 [+0.293, +0.424]	+0.007 [-0.011, +0.026]
Qwen3-14B-Base	+0.380 [+0.303, +0.467]	-0.023 [-0.055, +0.006]

F PER-LANGUAGE SUPERVISION COMPARISONS

These are descriptive per-language results underlying the source-of-supervision comparisons. The "NN English" column is retained as an exploratory per-language retrieval condition.

Table 10: Per-language RER scores for every condition, 0.5B.

held out	full	−English	English→35	size-matched	English only	NN English
de	0.287	0.205	0.219	0.188	0.175	0.138
es	0.297	0.217	0.217	0.183	0.178	0.355
fa	-0.160	-0.130	-0.133	-0.114	-0.093	-0.242
fr	0.254	0.184	0.195	0.178	0.251	0.266
ko	-0.018	-0.003	-0.003	-0.034	0.065	-0.094
nl	0.240	0.158	0.171	0.148	0.214	0.341
pl	0.023	0.020	0.021	0.010	0.053	0.101
pt	0.391	0.388	0.375	0.317	0.214	0.021
ru	0.176	0.159	0.152	0.109	0.032	-0.052
zh	0.201	0.136	0.178	0.146	0.293	0.296
macro	0.169	0.133	0.139	0.113	0.138	0.113

Table 11: Per-language RER scores for every condition, 1.5B.

held out	full	−English	English→35	size-matched	English only	NN English
de	0.400	0.388	0.392	0.371	0.309	0.308
es	0.550	0.528	0.531	0.489	0.470	0.354
fa	-0.016	0.018	0.009	-0.056	-0.048	-0.117
fr	0.502	0.463	0.470	0.436	0.424	0.437
ko	0.192	0.181	0.186	0.167	0.214	0.116
nl	0.543	0.512	0.513	0.470	0.371	0.299
pl	0.242	0.230	0.227	0.188	0.135	0.032
pt	0.495	0.481	0.475	0.421	0.322	0.186
ru	0.337	0.327	0.317	0.281	0.177	0.257
zh	0.479	0.414	0.441	0.408	0.418	0.648
macro	0.372	0.354	0.356	0.318	0.279	0.252

Table 12: Per-language RER scores for every condition, 7B.

held out	full	−English	English→35	size-matched	English only	NN English
de	0.435	0.454	0.454	0.451	0.413	0.426
es	0.630	0.619	0.621	0.599	0.593	0.479
fa	0.165	0.141	0.153	0.158	0.142	-0.190
fr	0.517	0.489	0.485	0.528	0.490	0.585
ko	0.330	0.330	0.327	0.299	0.329	0.282
nl	0.588	0.632	0.615	0.599	0.426	0.633
pl	0.403	0.389	0.390	0.430	0.386	0.258
pt	0.658	0.629	0.625	0.550	0.475	0.459
ru	0.482	0.520	0.514	0.457	0.334	0.484
zh	0.533	0.510	0.509	0.460	0.454	0.574
macro	0.474	0.471	0.469	0.453	0.404	0.399

Reverse family restriction. Restricting the eight primary Indo-European target languages to non-Indo-European supervision leaves 236 source terms. Across the four models examined, size-matched controls come within approximately 0.04 of the restricted scores, indicating that sample count explains at least part of the reduction. The comparison does not support a general estimate of linguistic-distance effects and is retained here as an exploratory diagnostic.

Table 13: Per-language RER scores for every condition, Llama-3.1-8B.

held out	full	−English	English→35	size-matched	English only	NN English
de	0.580	0.501	0.538	0.569	0.504	0.443
es	0.707	0.672	0.668	0.592	0.536	0.388
fa	0.220	0.227	0.214	0.214	0.160	0.059
fr	0.529	0.521	0.521	0.512	0.501	0.504
ko	0.346	0.392	0.369	0.359	0.389	0.222
nl	0.724	0.747	0.736	0.660	0.593	0.564
pl	0.494	0.478	0.480	0.449	0.362	0.040
pt	0.692	0.670	0.671	0.552	0.432	0.198
ru	0.435	0.489	0.474	0.415	0.254	0.243
zh	0.475	0.510	0.497	0.493	0.539	0.444
macro	0.520	0.521	0.517	0.482	0.427	0.310

Table 14: Per-language RER scores for every condition, Gemma-2-9B.

held out	full	−English	English→35	size-matched	English only	NN English
de	0.586	0.583	0.585	0.583	0.513	0.711
es	0.705	0.672	0.675	0.654	0.673	0.669
fa	0.190	0.240	0.230	0.226	0.304	0.306
fr	0.684	0.669	0.674	0.630	0.569	0.654
ko	0.405	0.389	0.388	0.338	0.355	0.366
nl	0.623	0.651	0.640	0.631	0.602	0.738
pl	0.552	0.566	0.554	0.556	0.481	0.511
pt	0.769	0.738	0.734	0.700	0.574	0.740
ru	0.423	0.460	0.452	0.439	0.393	0.520
zh	0.512	0.555	0.557	0.621	0.624	0.766
macro	0.545	0.552	0.549	0.538	0.509	0.598

Table 15: Per-language RER scores for every condition, Qwen3-14B-Base.

held out	full	−English	English→35	size-matched	English only	NN English
de	0.639	0.642	0.640	0.612	0.504	0.332
es	0.716	0.657	0.666	0.616	0.551	0.535
fa	0.313	0.308	0.307	0.286	0.147	0.065
fr	0.676	0.552	0.570	0.602	0.584	0.425
ko	0.414	0.424	0.427	0.419	0.417	0.254
nl	0.723	0.693	0.696	0.660	0.558	0.148
pl	0.456	0.373	0.393	0.454	0.371	0.329
pt	0.648	0.634	0.645	0.545	0.399	0.320
ru	0.626	0.658	0.653	0.588	0.300	0.461
zh	0.500	0.543	0.540	0.507	0.464	0.749
macro	0.571	0.548	0.554	0.529	0.429	0.362

G REGION-HOLDOUT DETAIL

The following tables retain the original target-region assignments. Each region contributes equally to the main region-macro result, regardless of the number of eligible language cells. Regions vary substantially in baseline performance and in sensitivity to deletion.

Table 16: Language-level nearest-source coverage in the primary ten held-out languages. For each test term, the nearest permissible training centroid is measured by Euclidean distance in OKLab. Entries are the fractions with a training neighbour within the stated threshold; the training pool includes all remaining validated languages.

held-out language	terms	≤ 0.02	≤ 0.04	≤ 0.06
German	56	0.929	1.000	1.000
Spanish	51	0.961	1.000	1.000
Persian	45	0.956	1.000	1.000
French	45	1.000	1.000	1.000
Korean	139	0.921	1.000	1.000
Dutch	22	1.000	1.000	1.000
Polish	23	1.000	1.000	1.000
Portuguese	28	1.000	1.000	1.000
Russian	35	0.886	1.000	1.000
Chinese	84	0.952	1.000	1.000

Table 17: Selection sensitivity for the region penalty. The fixed column reuses the full-coverage layer and penalty; the reselected column repeats both choices after removing the target region. Random removal is averaged over ten seeds. Intervals are paired descriptive resampling intervals over language–region cells, with random seeds sampled within cells.

Model	Random	Region fixed	Region reselected	Reselected Δ	95% interval
Qwen2.5-7B	.374	.157	.186	$-.188$	$[-.298, -.079]$
Llama-3.1-8B	.391	.194	.203	$-.187$	$[-.326, -.083]$

Table 18: Region holdout, per region. Each cell is the RER score on held-out-language terms inside that region, averaged over the confirmatory languages contributing terms to it. *full* is the unmodified training set, *out* removes the region from every language, *sm* removes the same number of terms drawn across all regions.

	Qwen2.5-7B			Gemma-2-9B			Qwen3-14B-Base		
region	full	out	sm	full	out	sm	full	out	sm
0	0.528	0.259	0.516	0.494	0.201	0.503	0.569	0.333	0.567
1	0.202	0.142	0.207	0.336	0.229	0.333	0.337	0.228	0.330
2	0.519	0.021	0.497	0.546	0.201	0.541	0.621	0.165	0.604
3	0.245	0.158	0.243	0.254	0.236	0.283	0.468	0.419	0.452
4	0.356	0.147	0.368	0.589	0.341	0.579	0.510	0.278	0.510
5	0.423	0.215	0.415	0.475	0.359	0.476	0.451	0.092	0.443

Training-size sensitivity. Existing runs for Qwen2.5-7B and Llama-3.1-8B repeat this comparison with initial pools of 46, 100, 200, 500, or all available training terms. The region-minus-random difference remains negative throughout: approximately -0.23 to -0.26 for Qwen and -0.21 to -0.24 for Llama. These runs average eligible language–region cells equally and use the sampled pool to define the shared constant baseline at each size. They are a sensitivity analysis, not directly interchangeable estimates of Table 3. They show persistence over the tested sizes, not elimination of all finite-sample or regularization effects.

Table 19: The six regions: k-means clusters of the 1160 human centroids in OKLab, k fixed at 6 before any model was run. *languages* counts confirmatory languages contributing at least five test terms to that region.

region	terms	languages	centroid L	centroid a	centroid b
0	217	9	0.649	+0.118	-0.013
1	164	5	0.490	+0.042	+0.024
2	287	10	0.805	-0.053	+0.064
3	223	7	0.653	-0.025	+0.022
4	120	6	0.358	+0.036	-0.060
5	149	6	0.534	+0.051	-0.116

Table 20: Subword tokens per colour term, by language and model, in the carrier used for extraction. Character coverage and span contiguity are 1.000 for every language and every model, so no term's representation is pooled over a truncated or discontiguous span.

language	Qwen2.5-7B	Llama-3.1-8B	Gemma-2-9B	Qwen3-14B-Base
ru	4.89	4.37	3.60	4.89
fa	4.71	3.53	3.38	4.71
zh	3.50	2.68	1.82	3.50
pl	3.39	3.26	2.65	3.39
ko	3.11	2.87	2.63	3.11
fi	2.92	2.92	2.85	2.92
de	2.84	2.82	1.95	2.84
es	2.69	2.69	1.76	2.69
nl	2.68	2.68	1.73	2.68
pt	2.68	2.68	1.75	2.68
en	2.28	2.28	2.01	2.28
fr	2.27	2.24	1.91	2.27
sv	2.21	2.21	1.89	2.21
ro	1.80	1.60	1.20	1.80

H TOKENIZATION AND EXPLORATORY LANGUAGE DIFFERENCES

Across ten evaluation languages, tokenizer fertility correlates negatively with the RER score (Spearman correlations approximately -0.57 to -0.77, with unadjusted $p < .05$ in four models). This analysis followed exploration of lexicon size, family, script, and orthographic overlap. It is an exploratory association, not a confirmatory mechanism or an explanation independent of language exposure. Correlated models and shared tokenizers also prevent treating model-wise correlations as independent replications.

I RAW REPRESENTATION DISTANCES AND HUMAN NAMING DISTANCES

Successful readout transfer leaves a complementary question: do distances in the original representations track distances between human naming centroids? We compare raw cosine distances with Euclidean OKLab centroid distances at the layer already selected for each outer probe fold. No layer is reselected using distance correlations. Human distances come from naming centroids rather than direct similarity judgments.

For each of the ten primary target languages, we retain its probe-evaluation lexicon and saved selected layer. Within-language correlations use all distinct pairs of these test words. Cross-language correlations pair the same target words with every word in each of the other nine primary evaluation languages. We compute a separate correlation for each source language, average the sources equally, and then average the target folds equally. These 90 directed blocks retain the original cross-language analysis population. English and the three small lexicons are excluded from this distance diagnostic, although they can contribute to probe training. The two directions of a language pair may use different selected layers.

We report ordinary Spearman correlations and a reliability-weighted sensitivity analysis. The latter computes the Pearson correlation of the same ordinary midranks, weighted by $w_i w_j$, where w_i and w_j are saved term-reliability weights normalized within each pair block. This weighting emphasizes

Table 21: Raw-distance correspondence at the layer already selected for each held-out-language probe. U denotes ordinary Spearman correlation; W denotes product-reliability-weighted Pearson correlation of ordinary midranks. Cross-language scores average sources within target folds and then targets equally. Within-language scores use the exact held-out lexicons. These coefficients are not on the RER scale.

	Cross-language		Within-language	
Model	U	W	U	W
Qwen2.5-0.5B	0.000	0.019	0.029	0.059
Qwen2.5-1.5B	0.032	0.087	0.048	0.100
Qwen2.5-7B	0.067	0.140	0.055	0.102
Llama-3.1-8B	0.095	0.159	0.094	0.164
Gemma-2-9B	0.103	0.148	0.084	0.106
Qwen3-14B-Base	0.115	0.198	0.097	0.148

Table 22: Direct correspondence between raw cosine distances and human naming-centroid distances (macro Spearman correlation). Cross-language results average 45 language pairs; within-language results average ten languages. No readout is fitted. Fixed-layer reference results; the main text reports probe-selected layers.

	Final layer		Middle layer	
Model	Cross-language	Within-language	Cross-language	Within-language
Qwen2.5-0.5B	0.000	-0.020	-0.007	0.015
Qwen2.5-1.5B	0.008	0.003	0.029	0.036
Qwen2.5-7B	0.008	-0.007	0.078	0.077
Llama-3.1-8B	0.063	0.019	0.082	0.079
Gemma-2-9B	0.050	0.028	0.132	0.118
Qwen3-14B-Base	0.030	0.023	0.095	0.073

pairs with reliable endpoints, but it is not identical to weighting single-term prediction errors. No learned projection or centring is applied. Because words and outer training folds overlap, pair counts are not independent replicates; the coefficients are descriptive.

At the probe-selected layers, cross-language correlations range from 0.000 to 0.115 without weighting and from 0.019 to 0.198 with reliability weighting (Table 21). All six unweighted selected-layer macro scores exceed their fixed-final-layer counterparts except Qwen2.5-0.5B. Weighting increases all six selected-layer scores. For example, Qwen3-14B-Base changes from 0.030 at its fixed final layer to 0.115 at fold-selected layers, and to 0.198 with weighting. Selected layers do not uniformly outperform fixed middle layers. Appendix J retains the fixed-layer analysis and matched weighting comparisons.

The results qualify the initial weak fixed-layer correspondence: raw-distance alignment depends on layer choice and target reliability weighting. Matching the probe layers reveals positive correspondence, especially in larger models, alongside successful cross-language prediction. A rank correlation and a prediction-error reduction remain different measurements, so their numerical sizes do not establish a decoding–geometry gap or show how much information a projection extracts. Product pair weights also differ from probe evaluation weights. These data support reporting both properties, not claiming that raw representations lack colour structure. Lexical-form similarity and the choice of cosine distance remain possible influences on the distance results.

J FIXED-LAYER DISTANCE SENSITIVITY

The initial diagnostic used the stored final layer and $\lfloor L/2 \rfloor$ (where L is the final stored layer index), with unweighted correlations across 45 unordered language pairs. These layers were specified before that computation. Its values are retained below; the main text now prioritizes the existing probe-selected layers. Recomputing fixed layers with 90 directed blocks reproduces the original unweighted macro values, because each pair contributes twice at the same layer.

For Qwen2.5-0.5B, Qwen2.5-1.5B, Qwen2.5-7B, Llama-3.1-8B, Gemma-2-9B, and Qwen3-14B-Base, respectively, product-reliability-weighted final-layer cross-language correlations are -0.012, -0.007, 0.006, 0.061, 0.052, and 0.028. Weighted middle-layer values are 0.023, 0.083, 0.145,

0.143, 0.186, and 0.166. Full directed-block outputs retain layer indices and both weighting conditions. These sensitivities do not constitute independent significance tests or establish that selected layers maximize geometric correspondence.

K SCOPE OF DISTRIBUTIONAL EVIDENCE

Participant-split analyses in ten sufficiently supported languages yielded median residual correlations of 0.640–0.773 after accounting for centroid distance and distribution radius. These languages include English and exclude Polish, so they are not identical to the primary evaluation set. The residual structure is reproducible under this construction, but within-language reliability alone does not establish language specificity.

Earlier project analyses tested whether model distances add held-out predictive value for human naming-distribution distances after controlling for centroid geometry, distribution width, and lexical features. These analyses used term-disjoint evaluation folds and human-derived nuisance features at evaluation time. They test conditional explanatory value rather than reconstruction without human information. The language-macro increment was approximately 0.00393 in Qwen2.5-7B and -0.00024 in Llama-3.1-8B, so the specified cross-architecture criterion was not met. We do not claim general recovery of full naming distributions from the centroid-decoding results. This distance-based non-replication also does not establish the absence of all language-specific information.