

Do Multilingual Models Represent Colour Differently Across Languages? The Gap Is in the Vocabularies

Rong Wang

University of Tübingen

rong.wang@uni-tuebingen.de

Abstract

Languages differ in how much of a word’s meaning is visible in its spelling. A colour term may name its own category in its written form or give no hint of it at all, and languages differ widely in how many of their terms do. This matters for a standard test. A model is said to represent colour well if it places two colour words close together whenever the colours they name look alike, and far apart when they do not. Comparing two languages means running that test on each language’s own colour words. But where the spelling already names the category, a model can place the words correctly by reading their spelling alone, so the test is easier in some languages than in others. We quantify this for four languages and two multilingual models, Qwen2.5-7B and Llama-3.1-8B. The proportion of terms whose written form names their category ranges from 0% to 92% across our four lexicons, and the proportion of the score that disappears once written-form similarity is removed tracks it closely ($r = 0.84$): 0–5% in Spanish and English, 32–71% in Chinese, 37–42% in German. On the raw test Chinese scores higher than English. Counting only what the model could not have got from spelling, English is ahead. We then test all four languages on the same 43 colours, so that no vocabulary gives more away than any other. Their scores become close enough that we cannot separate them at this sample size, though a finer comparison shows the four languages still arrange those colours somewhat differently. Much of what a conventional comparison reports then looks like a fact about how each language spells its colour words rather than about the model.

1 Introduction

Colour vocabularies are written very differently across languages. Chinese writes *pink* as 粉红 (*fěnhóng*, ‘powder red’) and *peach* as 桃红 (*táohóng*, ‘peach red’), both containing 红 (*hóng*), the character for red. German compounds in the same

way inside a single word: *Feuerrot* ‘fire red’ contains *rot*, *Signalgelb* ‘signal yellow’ contains *gelb*. English does not: *crimson* and *scarlet* are both reds, but nothing in their spelling says so. Spanish mostly does not either, because it modifies colour outside the word rather than inside it, writing *navy* as *azul marino* and *teal* as *azul cerceta*, two words each. How much of a colour’s identity is visible in the spelling of its name is therefore a property that varies sharply from language to language.

This matters for a standard kind of experiment. To ask whether a language model represents colour, we look at where it places each colour word inside itself. If *crimson* sits near *scarlet* and far from *turquoise*, the model has arranged those three words the way the colours themselves are arranged. Checking this over every pair of words in the vocabulary, and asking how consistently the arrangement matches, gives a single number. We call it the model’s *colour alignment score*: it is high when words the model places together name colours that do look alike. To ask whether a model represents colour better in one language than another, the natural move is to compute the score in each language, using that language’s colour words, and compare.

The two scores are not equally hard to earn. In Chinese, a model that noticed only that 粉红 (*fěnhóng*) and 桃红 (*táohóng*) share a character would already place them near each other, and it would be right. In English, nothing about the spelling of *crimson* and *scarlet* says they are both reds; a model can only place them together by having learned what they denote. The Chinese score can be earned partly from the writing system. The English score cannot.

We want to be precise about what this does and does not mean. A Chinese speaker also uses 红 (*hóng*) in 粉红 (*fěnhóng*) to know that the colour is a kind of red. That is not a shortcut but how the language works, and a model exploiting it is not cheating. The problem is comparison, not compe-

tence. When one language gives away part of the answer in the word and another does not, the same number means different things in the two cases, and a gap between them cannot be read as a fact about the model. Our question is how big that difference is, and what is left of a cross-lingual gap once it is removed.

RQ1. How much of a colour alignment score can be earned from the spelling of the colour words alone, and does that amount differ enough across languages to change which language scores higher?

RQ2. If the same colour concepts are used in every language, so that no language has an easier vocabulary, does a cross-lingual difference remain?

Neither question has a settled answer. If spelling turned out to contribute little, comparing scores across languages would be sound as it stands; if matching the concepts left the gaps intact, the vocabularies would not be the explanation for them.

We study English, German, Spanish and Chinese in two multilingual models, Qwen2.5-7B and Llama-3.1-8B. Studies of colour representation have been run in English, where the written form is a constant. Carrying the design to other languages turns that constant into a variable, and nothing in the usual procedure registers the change. We measure how much each vocabulary gives away in its spelling, from the written forms alone with no model involved. We then separate the part of a score that depends on spelling from the part that does not. Finally we remove the difference between the vocabularies altogether, by naming the same 43 colours in all four languages. The comparison does not give the same answer before and after.

2 Related Work

Colour representation from text. Colour is the standard case for asking whether a text-only model recovers structure that is fixed outside language, because the target is a measured coordinate rather than a judgement. Model representations of colour terms are compared against CIELAB coordinates by representational similarity analysis (Abdou et al., 2021; Kriegeskorte et al., 2008), or models are asked to map onto grounded conceptual spaces (Patel and Pavlick, 2022). The findings are mixed. The designs share a feature that matters here: they are run in English, where every colour term is spelled opaquely, so the written form is a constant and never has to be controlled.

Written form as a route to the answer. A writing system that encodes part of a semantic classification lets orthography stand in for meaning. Chinese characters do this for colour, and German compounding does it for colour without any logographic script. The possibility is easy to state, and we know of no measurement of it. Three things about it are open: how large the effect is, whether it requires a logographic script, and whether it is big enough to reverse a comparison. We answer all three. Up to 71% of a score is lost when written-form similarity is partialled out, a share of the measured score, not an estimate of how much spelling causally contributes. It does not require a logographic script, since the most transparent of our four vocabularies is an alphabetic one. And it does reverse a comparison.

Comparing representations across languages.

Running the same probe in several languages is a standard tool, used to ask whether a model builds one shared space or many (Conneau et al., 2020; Artetxe et al., 2020) and to compare the semantic structure of languages themselves (Thompson et al., 2020). Colour suits it, since languages divide the space differently within limits that recur across unrelated languages (Berlin and Kay, 1969; Kay and Regier, 2003; Gibson et al., 2017). But carrying a monolingual design across languages forces a choice the monolingual version never had to make: which words to use for each language. Taking each language’s own vocabulary is the natural answer and the only one that keeps the terms idiomatic; it also turns the written form from a constant into a variable. The nearest control for this is in bilingual lexicon induction, where words spelled alike *in the two languages* seed a dictionary and performance falls off for distant pairs (Søgaard et al., 2018). Our confound is the other kind: similarity between the written forms of a single language’s own words.

A complementary line asks the same question causally rather than correlationally. Maraia et al. (2026) construct steering directions for syllogistic reasoning in English and find that they transfer to other languages; Dumas et al. (2025) patch activations to change a concept while holding the output language fixed. Both argue that some internal structure is shared rather than rebuilt per language, as does Peng and Søgaard (2024) representationally. These results and ours bear on one issue from opposite ends: an intervention defined in one language carrying over to another argues for shared struc-

ture, while our measurements show that an apparent cross-lingual *difference* can be an artefact of the vocabularies compared. Surface-level cross-lingual scores are weak evidence about representational content in either direction.

Colour naming as a behavioural test. A parallel literature asks what models *do* with colour words rather than how they represent them. Mylonas et al. (2026) find that a model’s lexical translation accuracy and its colour correspondence come apart, and the word can be right while the colour it points to is wrong. Feldman and Peng (2026) find GPT-4o’s naming degrades near category boundaries and at low chroma. What both hold constant is the vocabulary: each language is asked about its own words. That is the step we examine here.

3 Method

The experiment runs in four steps, in this order: choose the words and fix their colour coordinates (§3.1); measure how much each vocabulary gives away in its spelling, with no model involved (§3.2); read representations out of the models (§3.3); compute the alignment score and split it into the part spelling can account for and the part it cannot (§3.4).

3.1 Colour stimuli

Each language contributes its own colour terms, taken from an established reference for that language rather than by translating an English list, since vocabularies differ in size and in which regions of the space they name densely (Regier et al., 2007; Gibson et al., 2017). English comes from the crowdsourced xkcd survey plus CSS4, German from RAL Classic with official CIE $L^*a^*b^*$ values, Chinese from a traditional colour dictionary, Spanish from the Spanish Wikipedia list. Every term carries one sRGB or Lab specification, which we convert to CIELAB under D65 so that all four lexicons live in one perceptual space. We keep single-word terms with chroma $C^* \geq 25$, which drops greys and near-greys whose hue is unstable. English is additionally restricted to terms carrying a fastText English vector, a frequency-like filter inherited from an earlier stage of this work and unrelated to spelling. This leaves 235 English, 106 Chinese, 59 German and 41 Spanish terms.

These four lists answer RQ1, where each language is scored on its own vocabulary. RQ2 needs the vocabularies not to differ at all, so it uses a

	source	N	final	head	ρ_{form}
Spanish	Wikipedia list	41	12%	0%	−0.038
English	xkcd + CSS4	235	18%	11%	−0.058
Chinese	traditional dictionary	106	79%	70%	−0.245
German	RAL Classic	59	90%	92%	−0.397

Table 1: Orthographic transparency of the four colour vocabularies, computed from the written forms alone with no model involved. **final**: shares its final token with some other term, i.e. the read position in a last-token analysis. **head**: the written form contains a basic colour term (*Feuerrot* contains *rot*; *crimson* does not contain *red*). ρ_{form} : Spearman correlation between a pair’s written-form overlap and its CIELAB distance. The more negative, the more the spelling alone already predicts the colour. Transparency spans 0–92%. German is alphabetic yet the most transparent, so the effect tracks morphology rather than script type; Spanish, which builds most of its colour vocabulary analytically (85 of the 163 entries in its reference list run to more than one word and are excluded here), anchors the opaque end. The ordering is monotonic across all three measures. $p < 10^{-20}$ except Spanish, whose ρ_{form} is not significant ($p = 0.3$), as expected for a vocabulary with no transparent terms.

separate set: the 43 colour concepts that have a one-to-one single-word term in all four languages. Each concept carries a single CIELAB coordinate that all four languages name, so the four scores are computed over one and the same set of colours.

3.2 How much a vocabulary gives away

Before any model is involved, we measure the written forms alone. For a language ℓ with terms $w_1 \dots w_n$, let F_{ij}^ℓ be the Jaccard overlap between the written forms of w_i and w_j : over characters for Chinese, over character trigrams for the alphabetic languages, since single letters carry no morpheme information and would make every pair look alike. If a vocabulary writes its categories into its words, pairs with high F^ℓ will tend to name nearby colours, so the correlation ρ_{form} between F^ℓ and CIELAB distance is negative. This is the quantity that later has to be controlled for, and it can be read off the lexicon on its own.

Two coarser measures need no matrix. One is the share of terms that share their final token with another term, which matters because that token is the read position (§3.3). The other is the share whose written form contains a basic colour term.

Table 1 shows that transparency spans almost the whole available range, and that the ordering is not the one a script-based account predicts. German

is alphabetic and the most transparent of the four. Spanish is the least, because it builds most of its colour vocabulary analytically: 85 of the 163 entries in its reference list run to more than one word (*azul celeste*), and are therefore excluded. German builds its modifiers inside the word instead. What varies is morphology, not script type.

3.3 Representations

We take residual-stream activations from Qwen2.5-7B and Llama-3.1-8B, read at the term’s final token. Each term is presented in a carrier that forces the colour reading and is byte-identical within a language, so that everything except the term itself is held constant: *the colour X*, *die Farbe X*, *el color X*, *这种颜色叫X* (*zhè zhǒng yánsè jiào X*, ‘this colour is called X’). This gives one vector x_i per term at every depth in the network. We index depth by transformer block: block 0 is the embedding, and block ℓ is the residual stream after the ℓ -th block has been applied. Qwen2.5-7B has 28 such blocks and Llama-3.1-8B has 32. Tables report block 14, near the middle of both networks, since a word’s lexical content is usually most directly readable there rather than at the input or at the output, where the model is committing to a next token. §4 repeats the analysis in Qwen2.5-7B from block 4 to block 28 and finds the same pattern throughout, so nothing turns on this choice.

3.4 The score and its two parts

Build D^ℓ , the matrix of distances between the x_i , and correlate its off-diagonal entries with the corresponding CIELAB distances. This Spearman correlation is the *colour alignment score*: it is high when word pairs that sit close together in the model also name colours that sit close together in perceptual space. It is the number a cross-lingual comparison compares.

Some pairs can be scored correctly without knowing anything about colour. Chinese 粉红 (*fěnhóng*) and 桃红 (*táohóng*) look alike and name similar colours, so placing them together is right even for a model that only compares spellings; English *crimson* and *scarlet* name similar colours but look nothing alike, so the same model would place them apart and be wrong. To count only what spelling could not have supplied, we recompute the correlation with F^ℓ partialled out. This asks how well the model arranges the words *among pairs whose spellings are equally alike*. What survives is the part of the score spelling cannot account for; the

	N	ρ raw	ρ ctrl	attrib.	share
<i>Qwen2.5-7B</i>					
Spanish	41	0.242	0.242	0.000	0%
English	235	0.217	0.212	0.005	2%
Chinese	106	0.291	0.199	0.092	32%
German	59	0.431	0.273	0.158	37%
<i>Llama-3.1-8B</i>					
Spanish	41	0.232	0.233	-0.001	0%
English	235	0.118	0.113	0.005	5%
Chinese	106	0.160	0.045	0.114	71%
German	59	0.371	0.217	0.155	42%

Table 2: How much of each colour alignment score comes from spelling. Partial Mantel between representational distance and CIELAB distance at block 14, before and after partialling out written-form overlap; **attrib.** is the difference and **share** its fraction of the raw score. Read uncontrolled, Chinese outranks English in both models; controlled, the ordering reverses. Spanish, whose vocabulary is 0% transparent, shows no attributable share at all. Across these eight cells the attributable share tracks the lexicon’s transparency in Table 1 at $r = 0.84$ ($p = 0.009$).

difference between the two numbers is the part it can, and we report that difference as a share of the original score.

Both entries of a pair contribute to many other pairs, so the usual sampling distribution for a correlation does not apply, and we compare each score against a null built by permuting the colour labels across terms. The permutation leaves both distance matrices exactly as they were and destroys only their correspondence, so it isolates the one thing being tested. Reported z is the distance from that null in its own standard deviations.

4 RQ1: How Much of the Score Comes from Spelling?

Table 2 gives the share of each score that comes from written form. It is 2% and 5% for English, 0% in both models for Spanish, 32% and 71% for Chinese, 37% and 42% for German. Across the eight cells the share correlates with lexicon transparency at $r = 0.84$ ($p = 0.009$). Uncontrolled, Chinese scores above English in both models (0.291 vs 0.217; 0.160 vs 0.118). Controlled, the ordering reverses in both (0.199 vs 0.212; 0.045 vs 0.113), and the controlled Chinese score in Llama-3.1-8B sits just above the permutation null ($z = 2.9$).

Figure 1 plots the eight cells. The points fall on a line rather than into two clumps. Transparency is therefore a quantity that varies by degree, which a study has to measure and report, and not a label

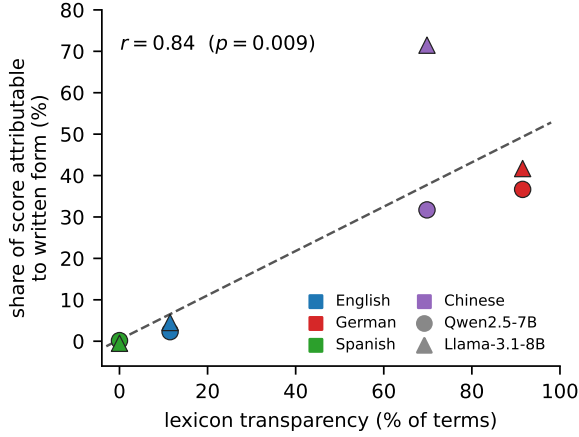


Figure 1: The share of a colour alignment score that comes from written form, against how much the lexicon gives away in spelling. Eight cells, four languages by two models; dashed line is the least-squares fit. The share rises with transparency across the whole range ($r = 0.84$, $p = 0.009$). The two opaque vocabularies, Spanish at 0% transparent and English at 11%, give up 0–5% of their score; the two transparent ones, Chinese at 70% and German at 92%, give up 32–71%.

that divides languages into two kinds.

Layer. Block 14 is one choice among many, so we repeat the analysis in Qwen2.5-7B at blocks 4, 9, 14, 20 and 28. Changing the block moves each language’s attributable share within a band: Spanish 0–2%, English 2–7%, German 30–40%, Chinese 25–77%. The two opaque vocabularies stay below 7% at every block and the two transparent ones stay above 25%, so the two sets of bands never meet. Which block is read changes how large the effect is, not which languages have one, and block 14 is representative rather than selected.

Not a suffix artefact. Many Chinese terms end in the nominalising 色 (*sè*) ‘colour’, a form shared across terms that says nothing about which colour. Removing it strengthens the coupling rather than weakening it. Excluding every term that contains it moves ρ_{form} from -0.245 to -0.350 , because what remains are compounds of colour morphemes themselves. On the model side the attributable share over those 82 terms is 31% for Qwen2.5-7B, essentially unchanged. For Llama-3.1-8B it is 43%, lower than 71%, so part of that figure was the suffix. It is still nine times the English share.

Not a paint-catalogue artefact. RAL Classic is an industrial standard, so the German figure could in principle belong to the catalogue rather than to German. The two are separable, because

RAL mixes ordinary German compounds a speaker would use unprompted (*Feuerrot*, *Zitronengelb*, *Kastanienbraun*) with coinages that exist only inside the standard (*Verkehrsrot*, *Signalgelb*, *Perlкупfer*). Dropping the second group leaves 44 of the 59 terms. Transparency does not fall with it: the share of terms containing a basic colour term is unchanged and ρ_{form} strengthens from -0.397 to -0.496 , since the industrial modifiers are the ones that say least about the colour they attach to. The attributable share does drop, from 37% to 30% in Qwen2.5-7B and 42% to 36% in Llama-3.1-8B, but stays several times English. German’s transparency is a property of how the language builds colour words, not of where we found them.

5 RQ2: Matching the Concepts Across Languages

In RQ1 each language was scored on its own word list, so the four scores were computed over four different sets of colours. A gap between them could mean two quite different things: that the model represents one language’s colour words better, or only that one vocabulary happened to name easier parts of the space. RQ2 removes the second reading. The 43 shared concepts carry one CIELAB coordinate each and are addressed four times, once per language in that language’s carrier, so all four languages are scored on exactly the same colours.

The design assumes translation equivalence, and it holds. If the four terms for a concept did not pick out the same referent there would be no shared target to score them against. Two things say they do. First, the language centroids sit close together next to the spread of the colours themselves: they have mean pairwise cosine 0.836 in Qwen2.5-7B and 0.746 in Llama-3.1-8B, while two colour words inside one language average only 0.620 and 0.649. Whatever separates the languages is small beside what separates the colours. Second, subtracting each language’s centroid leaves same-concept retrieval across languages at 39–43 of 43 and 34–41 of 43: given a term in one language, the nearest of another language’s terms is almost always its translation. The four languages sit in one concept space, up to a language-specific offset.

The answer to RQ2 is that the gap does not survive. Table 3 gives the form-controlled score for each language on the shared concepts. Written-form overlap is near zero throughout, as matched

	ρ_{form}	Qwen2.5-7B	Llama-3.1-8B
English	-0.035	0.221	0.099
German	-0.021	0.187	0.117
Spanish	-0.090	0.198	0.071
Chinese	-0.014	0.217	0.119

Table 3: Identical stimuli, four languages. Each of the 43 colour concepts carries one CIELAB coordinate and is addressed four times, once per language in that language’s carrier, so all four scores are computed over one and the same set of colours. Written-form overlap is near zero throughout (ρ_{form}), and the form-controlled alignment is then flat across languages, a spread of 0.034 in Qwen2.5-7B and 0.048 in Llama-3.1-8B, against within-model spreads of 0.214 and 0.253 across Table 2’s language-specific vocabularies. Chinese, which looked both best and worst there depending on whether form was controlled, is among the highest here in both models.

	En	De	Es	Zh
<i>Qwen2.5-7B</i>				
English	–	0.71	0.68	0.71
German	0.71	–	0.66	0.68
Spanish	0.68	0.66	–	0.57
Chinese	0.71	0.68	0.57	–
<i>Llama-3.1-8B</i>				
English	–	0.74	0.55	0.53
German	0.74	–	0.60	0.55
Spanish	0.55	0.60	–	0.36
Chinese	0.53	0.55	0.36	–

Table 4: Do two languages induce the same geometry? Spearman correlation between the representational distance matrices of each language pair, on the identical stimuli of Table 3. Against a reliability ceiling of 0.89 and 0.85 (same language, second carrier), these represent 75% and 65% of the reliable structure: the four languages largely share one colour space, but a quarter to a third of it remains language-specific even on identical stimuli.

concepts force it to be, so there is little left for the control to remove. The scores are then flat: 0.034 apart in Qwen2.5-7B and 0.048 in Llama-3.1-8B, against within-model spreads of 0.214 and 0.253 across the language-specific vocabularies of Table 2. On identical colours the four languages score alike. The RQ1 gaps were a fact about the vocabularies, not about the model. Chinese, which was both the best and the worst language in Table 2 depending on whether form was controlled, is among the highest here in both models.

Figure 2 shows this as a layout rather than as a number, with both models. Every decoded panel reproduces the target’s arrangement: greens to the left, warm colours to the upper right, blues and pur-

ples below. Every panel also pulls the arrangement in toward the origin, the most saturated quarter of the concepts reaching 55% and 56% of its target radius. Since that shrinkage is uniform across all eight panels and is what a regularised decoder does to any target, the comparison the figure supports is between panels rather than between a panel and the target.

Read that way, the panels within a row are close but not identical. Two languages place the same concept a median of 15.4 a^*b^* units apart in Qwen2.5-7B and 21.8 apart in Llama-3.1-8B, against a median target radius of 51.7, and the sign of both a^* and b^* survives for 65–70% and 53–63% of concepts in every language. The two models place the same concept in the same language 17.1 apart, which falls between the two within-model figures. Changing the language moves a colour about as far as changing the model does, and the four languages are not the larger source of the two.

Equal scores are not the same as one shared space. Two languages can each align with CIELAB equally well and still arrange the 43 colours differently relative to one another, so Table 4 asks that question directly, correlating the representational distance matrix of each language pair. The values are 0.57–0.71 and 0.36–0.74, lowest for Spanish–Chinese in both models. To know what they are worth we re-extracted every language under a second, unrelated carrier: a language correlates with itself at 0.89 and 0.85 across the two carriers, which is the most any pair could reach given measurement noise. The observed values are 75% and 65% of that. Concept identity transfers between these languages almost completely; the colour geometry built over it transfers by about two thirds.

6 Discussion

Our findings demonstrate that cross-lingual gaps in colour alignment scores, as conventionally measured, are largely artifacts of surface vocabulary rather than true differences in how language models represent colour concepts. When vocabularies are evaluated in isolation, orthographic transparency heavily inflates alignment scores ($r = 0.84$), accounting for up to 71% of the performance in morphologically explicit languages like Chinese and German.

A critical nuance revealed by our analysis is that this inflation is driven by compositional morphol-

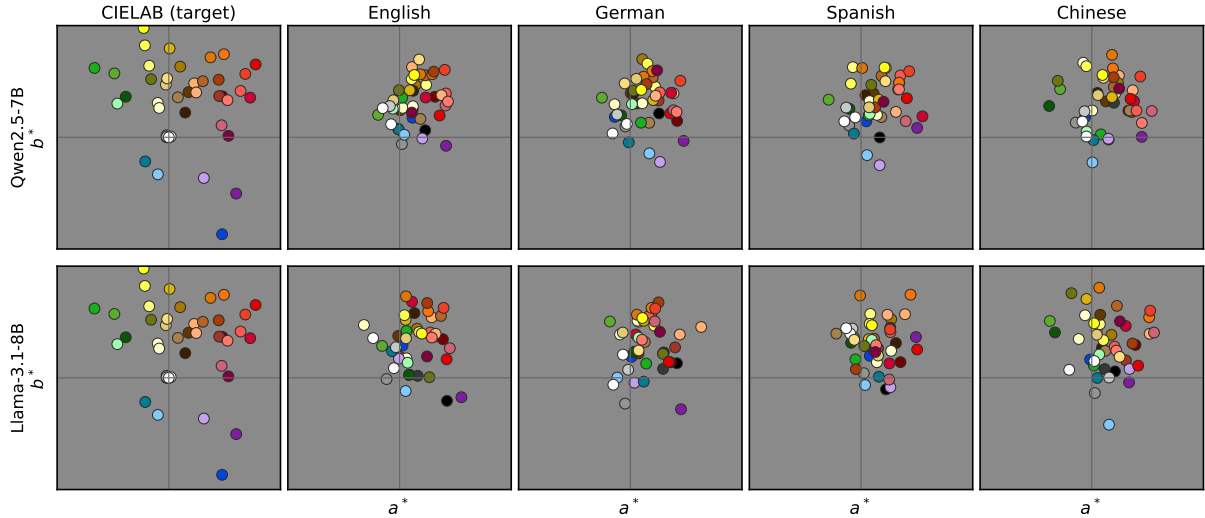


Figure 2: Identical stimuli, decoded. Each point is one of the 43 shared concepts, placed where a leave-one-out ridge decodes it in the a^*b^* plane and filled with the colour that concept actually denotes; block 14. The first column is the CIELAB target, the rest are the four languages, Qwen2.5-7B above and Llama-3.1-8B below. Every decoded panel keeps the target’s arrangement, greens to the left, warm colours upper right, blues and purples below, and shrinks it toward the origin: the most saturated quarter of the concepts reaches 55% and 56% of its target radius. That shrinkage is the same in all eight panels and is what a regularised decoder does to any target, so what the figure is for is the comparison between panels.

ogy rather than writing systems per se. German, an alphabetic language with internal compounding, shows an attributable share (30–42%) comparable to Chinese (32–71%), whereas Spanish and English show negligible surface effects (0 – –5%). This explains why monolingual studies in English could safely ignore orthographic leakage, while cross-lingual comparisons fail if surface transparency is not controlled.

Fixing the stimuli across languages eliminates the artificial gap: when forced to evaluate the identical 43 colour concepts, all four languages achieve remarkably flat alignment scores. This establishes a strict methodological requirement for future multilingual probing: evaluating models on language-specific, unconstrained lexicons yields ambiguous scores that conflate lexical naming conventions with representational structure. Valid cross-lingual comparisons require matched conceptual stimuli.

Importantly, eliminating vocabulary artifacts does not render cross-lingual representations identical. On matched items, 65–75% of the reliable colour geometry transfers across languages, leaving a persistent language-specific residual of one-quarter to one-third. Stripping away this surface-level noise isolates genuine geometric divergence, offering a finer-grained lens for multilingual semantic alignment in LLMs.

7 Conclusion

We asked what is being measured when a multilingual model’s colour representations are compared across languages. Much of the answer is the writing system. The share of a colour alignment score that does not survive partialling out written-form similarity runs from 0% to 71% over four languages and two models, tracks how far each vocabulary writes its categories into its words, and is large enough to reverse which language looks better represented. Naming the same 43 colours in every language removes almost all of that difference and leaves a language-specific residual of a quarter to a third. A comparison run on each language’s own vocabulary should be read as a statement about the vocabularies unless the stimuli were matched. We release the four annotated lexicons and the matched-concept set so that a comparison can match them.

Limitations

The scope is two models, four languages, one domain and one readout position. The shared-stimulus set is small ($N = 43$) to ensure genuine cross-lingual correspondence rather than translation, excluding the terms where languages diverge most: Chinese 青 (*qīng*) spans blue and green and has no English counterpart. Matching the concepts also changes colour coverage, vocabulary size and

composition, so it isolates the written form only to the extent those move together, and the residual we attribute to language-specific geometry is what survives our controls rather than what they have shown to be semantic.

Two restrictions bound the conclusion. We test colour alone, though nothing in the mechanism is specific to it; and our four vocabularies cover compounding inside the word (German), modification outside it (Spanish), opaque alphabetic (English) and logographic (Chinese), but not agglutinative or non-concatenative morphology. Both extensions are concrete rather than open-ended: the transparency score needs only written forms, so a second domain requires a lexicon with a semantic reference, and a morphologically distant language requires one whose category marker is an affix rather than a free morpheme, where subword tokenisation may not isolate it.

Acknowledgements

This research was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project ID 527671319 (“Komplexe Wörter im Kontext”).

References

- Mostafa Abdou, Artur Kulmizev, Daniel Hershcovich, Stella Frank, Ellie Pavlick, and Anders Søgaard. 2021. Can language models encode perceptual structure without grounding? a case study in color. In *Proceedings of the 25th Conference on Computational Natural Language Learning*.
- Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637.
- Brent Berlin and Paul Kay. 1969. *Basic Color Terms: Their Universality and Evolution*. University of California Press.
- Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6022–6034.
- Clément Dumas, Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2025. [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)*, pages 31822–31841.
- Anna Feldman and Jing Peng. 2026. [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)*, pages 136–147.
- Edward Gibson, Richard Futrell, Julian Jara-Ettinger, Kyle Mahowald, Leon Bergen, Sivalogeswaran Ratnasingam, Mitchell Gibson, Steven T. Piantadosi, and Bevil R. Conway. 2017. Color naming across languages reflects color use. *Proceedings of the National Academy of Sciences*, 114(40).
- Paul Kay and Terry Regier. 2003. Resolving the question of color naming universals. *Proceedings of the National Academy of Sciences*, 100(15).
- Nikolaus Kriegeskorte, Marieke Mur, and Peter Bandettini. 2008. Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2.
- Gabriele Maraia, Leonardo Ranaldi, Marco Valentino, and Fabio Massimo Zanzotto. 2026. [Can activation steering generalize across languages? a study on syllogistic reasoning in language models](#). In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2739–2753.
- Dimitris Mylonas, Rafique Ahmed, Akvile Sinkeviciute, and Alexandros Koliousis. 2026. [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*.
- Roma Patel and Ellie Pavlick. 2022. Mapping language models to grounded conceptual spaces. In *International Conference on Learning Representations*.
- Qiwei Peng and Anders Søgaard. 2024. [Concept space alignment in multilingual LLMs](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5511–5526.
- Terry Regier, Paul Kay, and Naveen Khetarpal. 2007. Color naming reflects optimal partitions of color space. *Proceedings of the National Academy of Sciences*, 104(4).
- Anders Søgaard, Sebastian Ruder, and Ivan Vulić. 2018. On the limitations of unsupervised bilingual dictionary induction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 778–788.
- Bill Thompson, Seán G. Roberts, and Gary Lupyan. 2020. Cultural influences on word meanings revealed through large-scale semantic alignment. *Nature Human Behaviour*, 4(10):1029–1038.