Lexical Availability and Human Distributional Agreement in GPT-4o's Color Naming

Anna Feldman
Montclair State University
feldmana@montclair.edu

Jing Peng
Montclair State University
pengj@montclair.edu

Abstract

We evaluate GPT-4o's color naming across nine languages using both synthetic and human-derived stimuli. Using hue wheels, fixed basic categories, low-chroma hue lines, and dense binned CIELAB grids, we separate lexical availability of color terms from distributional agreement with human color naming. GPT-4o reliably names vivid, high-chroma colors and reproduces several known language-specific distinctions under constrained settings. However, its performance degrades sharply for low-chroma colors and for stimuli near human category boundaries. In these regions, model–human divergence remains high. Overall, GPT-4o shows strong cross-linguistic lexical knowledge but does not reliably match human color-naming distributions, especially in low-chroma and boundary regions.

1 Introduction

Color naming is a classic testbed for theories of semantic structure, linguistic relativity, and perceptual organization. Languages partition color space differently, but these systems follow regular patterns shaped by universal perceptual constraints and language-specific lexical conventions (Berlin and Kay, 1969; Kay and Regier, 2003; Regier et al., 2007; Shepard, 1992; Jameson and D'Andrade, 1997; Zaslavsky et al., 2018). The Russian contrast between *goluboy* 'light blue' and *siniy* 'dark blue' illustrates how lexical boundaries affect discrimination and categorical perception (Winawer et al., 2007; Thierry et al., 2009).

Color naming lets us separate two behavioral questions that are often conflated. First, does GPT-4o have lexical availability of color terms in a target language, meaning that it can produce plausible language-specific color names for a given patch? Second, do its sampled naming distributions show distributional agreement with human color naming, meaning that the model assigns probability mass to color terms in ways similar to human annotators for the same stimuli? We use the second criterion as behavioral evidence about perceptual grounding, but we do not directly inspect the model's internal perceptual representations. This distinction echoes classic grounding arguments: symbolic competence does not guarantee perceptual categories (Harnad, 1990).

Multimodal language models (MLLMs) perform well on recognition, captioning, and open-ended reasoning (Li et al., 2022; Alayrac et al., 2022; OpenAI, 2023, 2024), but it is unclear how much perceptual structure they actually encode. Prior work shows that models name vivid colors reasonably well yet diverge from human behavior on desaturated or boundary-region stimuli and in languages with distinct lexical systems (Liang et al., 2025). Most existing evaluations focus on a single language, small synthetic palettes, or coarse accuracy, leaving open whether a model like GPT-4o matches human color-naming distributions across languages and across both synthetic and human-derived color spaces.

We ask whether a strong multimodal model's color naming is merely lexically plausible, or whether it also resembles human naming distributions across perceptually difficult regions of color space. This distinction matters because a model can know that Russian has both голубой and синий, or that English has terms such as teal and turquoise, without using those terms in human-like ways for ambiguous or low-chroma stimuli.

Our approach. We evaluate GPT-4o in four settings (Table 1) that progressively move from lexical probing to human-distribution comparison: open-vocabulary naming of synthetic high-chroma hues, fixed-vocabulary naming of the same hues, naming of human low-chroma hue-line stimuli, and naming of full binned CIELAB grids with human reference distributions (Section 4).

*Proceedings of the 15th Joint Conference on Lexical and Computational Semantics (*SEM 2026)*, pages 136–147

Main finding. GPT-4o names vivid colors well and recovers some known language contrasts. It fails on low-chroma and boundary regions, where it often replaces human chromatic terms with gray or beige. Divergence from humans stays high across languages. This shows strong lexical availability, but weak distributional agreement with human color naming in low-chroma and boundary regions.

We make **four contributions:**

1. We provide a multilingual behavioral evaluation of GPT-4o color naming across synthetic high-chroma hue wheels, fixed-vocabulary category tasks, human low-chroma hue-line stimuli, and full CIELAB grids with human naming distributions.
2. We separate *lexical availability* from *human-distribution agreement* by comparing open-vocabulary prompts, fixed-vocabulary prompts, and human-matched distributional evaluations across the same color spaces.
3. We identify recurring cross-linguistic failure patterns in low-chroma and boundary regions, including collapse to desaturated labels, reduced coverage of human color terms, vocabulary compression, and a strong association between human category concentration and model–human agreement.
4. We introduce control experiments (text-only ablation, saturation ratings, and low-chroma temperature sweeps) that help distinguish the effects of lexical priors, chroma sensitivity, and sampling behavior.

2 Related Work

2.1 Color naming across languages

Research since Berlin and Kay (1969) shows that languages partition color space in ways shaped by perceptual structure and communicative pressures (Kay and Regier, 2003; Regier et al., 2007; Zaslavsky et al., 2018). Russian has played a central role because the lexical split between *goluboy* 'light blue' and *siniy* 'dark blue' correlates with differences in discrimination, categorical perception, and memory (Winawer et al., 2007; Thierry et al., 2009). Work in perceptual psychology emphasizes that human color categories emerge from the interaction of universal perceptual constraints and language-specific lexical histories (Shepard, 1992; Jameson and D'Andrade, 1997). Color naming is therefore a robust benchmark for semantic theories and a sensitive test of perceptual grounding.

Neural models of grounded color reference show that pragmatic reasoning mainly improves performance on the hardest discriminations (Monroe et al., 2017). This reinforces the idea that fine-grained color distinctions expose the role of perceptual evidence. A longstanding concern in vision and language research is that strong textual priors can mask weak perceptual grounding (Cadène et al., 2019; Bisk et al., 2020). Our work addresses this issue directly by comparing model predictions to human distributions in low-chroma and cross-boundary regions where visual cues should dominate. The new text-only baseline and saturation probes further isolate how much structure GPT-4o derives from vision rather than from lexical frequency.

2.2 Human color naming datasets

The UW Color Naming corpus includes many more languages, but for this study we analyze nine: English, Russian, Chinese, Korean, German, French, Spanish, Polish, and Portuguese. After filtering (section 3), six languages (English, Chinese, Korean, German, Russian, and Spanish) retain dense CIELAB coverage and serve as our primary evaluation set. French, Polish, and Portuguese retain moderate coverage and are treated as exploratory.

2.3 Multilingual color naming and cross-lingual comparison

The Kim et al. (2019) dataset has supported work on salient colors, alignment across languages, and category structure. Most analyses focus on human cross-linguistic patterns rather than on model evaluation. Our study differs in purpose: we use these human naming distributions as the reference for evaluating a general-purpose multimodal model. This makes it possible to test whether a model trained on large-scale vision and language data matches human color-naming distributions, human uncertainty patterns, and cross-linguistic lexical boundaries.

2.4 LLMs and perceptual grounding

Multimodal language models achieve strong results on captioning, recognition, and high-level reasoning (Li et al., 2022; Alayrac et al., 2022;

OpenAI, 2023), but their perceptual grounding is usually evaluated through tasks that do not require fine-grained visual discrimination. A growing body of work argues that models often rely on strong linguistic priors rather than visual evidence (Cadène et al., 2019). Grounded semantics requires sensitivity to perceptual structure, not textual co-occurrence (Bisk et al., 2020).

Work on color naming indicates that models label vivid, high-chroma colors reasonably well but diverge from human judgments on desaturated and boundary-region stimuli and outside English (Liang et al., 2025). These evaluations typically rely on accuracy over small synthetic sRGB palettes and rarely examine distributional alignment with human data.

To our knowledge, no prior study evaluates a model like GPT-4o on synthetic hue wheels, human low-chroma hue lines, and dense CIELAB grids, nor compares model and human naming distributions across multiple languages.

Recent work in the audio domain reports a closely related dissociation. Chen et al. (2026) show that large audio-language models often rely on lexical or transcript-level cues rather than acoustic information when performing emotion and paralinguistic judgments. Despite strong task performance, these models fail to exhibit human-like sensitivity to fine-grained acoustic structure. Together with our results, this suggests that the gap between lexical competence and perceptual grounding may be a general property of current multimodal models, not limited to vision.

3 Methods

3.1 Overview

Our goal is to evaluate how GPT-4o names colors across multiple languages and to compare its naming distributions directly to human judgments. We run four complementary experiments that vary in stimulus type, response format, and evaluation metric: A) lexical range under minimal constraints, B) coarse category boundaries under fixed vocabularies, C) performance on perceptually ambiguous low-chroma stimuli, and D) distributional agreement with human naming across the full color space.

Experiments A and B probe the model's intrinsic lexical behavior in the absence of human data. Experiments C and D directly compare GPT-4o to human naming distributions and test whether its responses show human-like distributional structure rather than only lexical plausibility.

3.2 Languages

Experiments A–C use English and Russian, following prior work that emphasizes the English *blue* category and the Russian *синий* versus *голубой* distinction. Experiment D extends the analysis to nine languages selected from the UW Color Names corpus: English, Russian, Chinese, Korean, German, French, Spanish, Polish, and Portuguese.

Human coverage varies substantially across languages (see bin counts in Table 5). After filtering, English, Chinese, Korean, German, Russian, and Spanish retain the largest number of usable LAB bins. French, Polish, and Portuguese retain fewer bins and are included as exploratory languages rather than primary evaluation languages. The prompting pipeline, normalization procedures, filtering, and Jensen–Shannon divergence computations are identical across languages.

3.3 Stimuli

We use three classes of stimuli.

Synthetic hue wheels. We generate evenly spaced high-chroma HSV values (36 bins), convert them to sRGB, and render each as a flat color patch at a fixed resolution. All synthetic images are produced by the same script and passed to GPT-4o without further compression. These stimuli allow controlled probing of lexical behavior that is free from human dataset biases.

Human hue-line stimuli. We use the low-chroma hue lines from Kim et al. (2019). These are diagnostic for perceptual category boundaries because humans disagree more in low-chroma regions. Russian and English hue-lines are used.

Full CIELAB grids. Following Kim et al. (2019), we analyze human color-name judgments after aggregation in CIE L*a*b* (CIELAB) space. The UW corpus bins the full-color data into $10 \times 10 \times 10$ CIELAB bins and provides human naming distributions over surface color terms. All nine selected languages are evaluated in Experiment D. Each retained CIELAB bin is converted to sRGB and rendered as a uniform color patch before being shown to GPT-4o.

3.4 Binning of Human and Model Data

We quantize CIELAB by rounding each channel after dividing by 10: $\text{binL} = \text{round}(L/10)$, $\text{binA} =$

Exp.	Stimuli	Languages	Main question
A	Synthetic high-chroma hue wheel, open-vocabulary naming	EN, RU	What lexical range and language-specific distinctions emerge under response-constrained open-vocabulary prompting?
B	Synthetic hue wheel, fixed basic terms	EN, RU	Does GPT-4o place major category boundaries, such as the Russian голубой vs. синий distinction, in the expected regions?
C	Human low-chroma hue lines	EN, RU	Does GPT-4o match human color-naming distributions on perceptually ambiguous and diagnostic low-chroma stimuli?
D	Full binned CIELAB grid	EN, RU, ZH, KO, DE, FR, ES, PL, PT	How closely do GPT-4o's naming distributions match human data across the entire CIELAB space, including low-chroma and cross-boundary regions?

Table 1: Overview of the four experiments. EN = English, RU = Russian, etc.

round$(a/10)$, binB = round$(b/10)$. This groups nearby colors while preserving local structure in LAB space and ensures sufficient human annotation counts per bin for stable distributional comparisons.

3.5 Model Query Procedure

Each stimulus is rendered as a flat RGB image and passed to GPT-4o using the multimodal API. All experiments used the OpenAI GPT-4o model (identifier: gpt-4o) accessed through the OpenAI Python SDK (openai v2.9.0). All images use the same rendering pipeline as described in the Stimuli section.

Prompts. We use language-specific prompts requesting one natural color word or very short phrase in the target language only, with no explanation or additional commentary. The prompts are provided in Appendix A.

Sampling settings. Sampling temperatures are chosen to trade off lexical diversity and stability. Higher temperatures encourage exploration of the model's vocabulary, while lower temperatures enforce more deterministic category choices. We perform a targeted temperature sweep for low-chroma bins (Experiment C) and use fixed temperatures for the main experiments.

- Experiment A: open-vocabulary naming, 100 samples per hue, $T = 0.3$ and $T = 0.9$ (lexical exploration).
- Experiment B: fixed vocabulary, 20 samples per hue, $T = 0.1$ (sharp category decisions).
- Experiment C: hue-line, 20 samples per chip, $T = 0.7$ (moderate diversity).
- Experiment D: CIELAB grid, 3–5 samples per bin, $T = 0.7$.

3.6 Control Experiments

In addition to Experiments A-D, we run three small control experiments to clarify the role of linguistic priors, perceptual input, and sampling temperature. A text-only baseline reuses the hue-line prompts without images to estimate purely linguistic color-name priors. A saturation sanity check asks GPT-4o to rate the saturation of a few high- vs. low-chroma tiles, and a low-chroma temperature sweep probes whether the failure patterns depend on sampling temperature. Summary results are reported in Section 4.5.

3.7 Normalization

Model outputs are normalized by lowercasing, removing punctuation, stripping diacritics when appropriate, and performing language-specific corrections such as Russian *ë* to *e* and German *ß* to *ss*. We then apply language-specific head–modifier parsing to extract heads that correspond to surface-level human lexical entries.

3.8 Human Naming Distributions

Here, $p(t \mid V)$ (ptV) denotes the normalized probability that human annotators assign surface color term t to visual stimulus V.

For each LAB bin (L, A, B) we sum ptV values per surface color term to obtain a probability distribution,

$$p_{\text{human}}(t \mid L, A, B),$$

which we use throughout as the human naming distribution for that bin.

3.9 Surface and Head-Level Categories

Because human labels are highly variable, we compute: (1) **surface distributions**: raw normalized labels, and (2) **head distributions**: labels with modifiers stripped (e.g. *ярко-синий* → *синий*, *light blue* → *blue*). A small set of seed modifiers per language (for example *light, dark, bright* in English and *ярко-, светло-* in Russian) is augmented with additional modifiers automatically extracted from the human vocabulary. This procedure yields a compact set of heads while preserving the main lexical contrasts.

Examples of head–modifier pairs for several languages are listed in Table 10 in the Appendix. For

Russian, head categories are further collapsed using robust stems (for example *голуб*, *син*, *зелен*, *роз*, *фиол*, *сер*). For Chinese and Korean, head extraction relies on a small set of character-level roots supplied by the dataset rather than full morphological analysis. This approximation keeps heads compact but may merge genuinely distinct lexical items in some cases.

3.10 Filtering

Filtering is applied only to Experiment D, where human responses are aggregated into CIELAB bins. A bin is retained only if it has sufficient human annotation support, a sufficiently concentrated dominant human head category, and valid RGB values. Default thresholds are $k = 8$ annotations and $p_{\min} = 0.20$ dominant-head proportion, with slightly looser thresholds for languages with lower retained coverage or annotation density (e.g., Russian, Polish, and Portuguese use $k = 4$ and $p_{\min} = 0.15$). After filtering, the per-language retained bin counts are shown in Table 2. The released configuration also caps retained bins by language to control API cost.

Lang	EN	ZH	KO	ES	DE	RU	FR	PL	PT
Bins	150	150	150	92	48	80	27	36	50

Table 2: Retained CIELAB bins per language in Experiment D after applying frequency and proportion filters.

3.11 Evaluation Metrics

Jensen–Shannon divergence. We compute JS divergence between human and model head distributions. Because human–human JS values are not available and model distributions are sparsely sampled, we treat absolute magnitudes cautiously and focus on relative patterns across languages, chroma levels, and error types.

Top-1 agreement. We compare human and model top categories per bin.

Vocabulary compression. We measure

$$\text{vocab\_ratio} = \frac{|\{t : p_{\text{model}}(t \mid b) > 0\}|}{|\{t : p_{\text{human}}(t \mid b) > 0\}|},$$

the ratio of the number of distinct labels used by the model to the number used by humans in each bin, where t ranges over color terms and b denotes a CIELAB bin. This measure ignores label overlap and may exceed 1 if the model produces more distinct labels than humans. Unlike the in-vocabulary mass in Table 5, which captures probability overlap, vocab_ratio is a count-based measure of inventory size used to characterize collapse or expansion behavior across bins.

Human category concentration. We compute the proportion of human responses assigned to the dominant head category per bin and test whether stronger human category agreement predicts lower model–human divergence.

Automatic error typing. For descriptive analysis, bins are grouped into broad quality categories (good, mid, bad) based on head-level JS divergence thresholds.

Per-hue vocabulary size (Experiment A). For the open-vocabulary hue-wheel experiment, we also measure the number of distinct labels GPT-4o uses for each hue and language. For hue index h and language ℓ, we define

$$\text{vocab\_size}(h, \ell) = |\{t : p_{\text{model}}(t \mid h, \ell) > 0\}|, \quad (1)$$

that is, the number of distinct normalized color terms observed across the n samples for that hue. Large values indicate that the model does not settle on a stable category for that color and instead spreads probability mass over many near-synonymous labels.

4 Results

We organize results by experiment. Experiments A–B probe GPT-4o's lexical behavior on synthetic hue wheels. Experiment C evaluates the model on human low-chroma hue-lines. Experiment D compares human and model naming distributions across the full binned CIELAB grid in nine languages.

Experiments A and B primarily evaluate lexical availability and category boundary behavior under controlled prompting, whereas Experiments C and D evaluate distributional agreement with human naming.

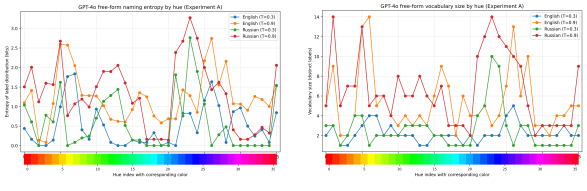

Figure 1: Entropy (left) and vocabulary size (right) by hue index for English and Russian under two sampling temperatures ($T = 0.3$ and $T = 0.9$).

4.1 Experiment A: Open-vocabulary Naming on Synthetic Hue Wheels

In this experiment, we collect 100 samples at $T = 0.9$ for each of the 36 hue bins. Figure 1 additionally compares entropy and vocabulary size under a lower sampling temperature ($T = 0.3$) to illustrate the effect of sampling diversity on lexical variability.

The left plot in Figure 1 shows that GPT-4o's naming entropy is lowest near prototypical reds and yellows and highest in the turquoise-blue-purple region, indicating stable lexical choices for warm hues and multiple competing labels for mid-spectrum hues. To complement entropy, we also examine the per-hue vocabulary size (1). The right plot in Figure 1 shows the vocabulary size by hue index for English and Russian. High-chroma reds and yellows typically elicit only one or two consistent labels, whereas the turquoise-blue–purple range often produces 8-15 distinct terms across the same number of samples. Russian vocabulary sizes are generally larger than English in this region, reflecting productive roots such as *голуб-*, *син-*, *бирюз-*, and *фиол-*.

Hue Range	EN top label	RU top label
0–1	scarlet, vermilion	алый
2–4	tangerine, amber	мандариновый, охра
5–7	sunflower, neon yellow	ярко-жёлтый, лимонный
8–12	lime, neon green	лаймовый
13–15	neon green	салатовый
16–17	mint, aqua	мятный, бирюзовый
18	cyan, sky blue, azure	бирюзовый
19–20	cyan, sky blue, azure	голубой
21	cyan, sky blue, azure	ярко-голубой
22–24	cobalt	ярко-синий, синий электрик, кобальтовый
25	electric blue	индиго
26–27	violet	фиолетовый
28–34	magenta, fuchsia	фуксия, малиновый
35	crimson	ализариновый

Table 3: Top open-vocabulary labels across synthetic hue ranges. Rows are ordered by increasing hue index around the HSV wheel. Split rows indicate changes in dominant labels within previously grouped ranges.

The top open-vocabulary labels per hue in Table 3 differ across languages: English often uses *teal*, *turquoise*, and *purple*, while Russian uses *бирюзовый*, *голубой*, *синий*, and *фиолетовый*. These cross-linguistic differences reflect lexical conventions rather than human-like distributional behavior, consistent with Experiment A probing lexical knowledge rather than grounded categories. Experiment A therefore establishes GPT-4o's lexical competence but does not test distributional agreement with human color naming.

4.2 Experiment B: Fixed Basic Categories on Synthetic Hue Wheels

In Experiment B, we restrict GPT-4o to basic color terms and lower the sampling temperature to $T = 0.1$ to isolate major category boundaries. Table 4 shows that Russian exhibits finer internal partitioning of the blue region under fixed-vocabulary prompting. These partitions approximate the major human boundaries but do not match them precisely. Entropy under fixed categories is near zero for most English hues, while Russian shows entropy peaks near boundary regions. GPT-4o therefore reproduces some language-specific lexical boundary behavior under constrained prompting. However, agreement under fixed vocabularies does not imply full distributional agreement with human color naming, as the low-chroma and full-grid results show.

These results confirm that GPT-4o can reproduce major lexical boundary behavior under constrained prompting, without demonstrating full distributional agreement with human color naming.

Boundary	EN	RU
RED → ORANGE	1–2	1–2
ORANGE → YELLOW	3–4	3–5
YELLOW → GREEN	6–7	6–7
GREEN → BLUE	16–18	17–20
BLUE → PURPLE	22–25	23–26
PURPLE → RED	33–35	33–35

Table 4: Experiment B: Category boundaries (hue ranges) under fixed basic color inventories.

4.3 Experiment C: Human Low-Chroma Hue-Lines

The low-chroma hue-line stimuli from Kim et al. (2019) are designed to probe perceptual ambiguity. We query GPT-4o (20 samples per chip, $T = 0.7$) and compare its head distributions to human naming. Figure 2 shows Russian results: JS divergence is high across nearly all bins, and English exhibits a similar pattern (figure omitted for space). GPT-4o frequently replaces diverse human chromatic responses with a narrow set of desaturated labels such as *серый*, *бежевый*, *gray*, and *beige*. The model rarely selects the human top category, especially in low-chroma boundary regions.

Because Experiment D generalizes this behavior across the full CIELAB space, we treat Experiment C as a diagnostic probe. We interpret this as evidence that GPT-4o appears to rely more heavily on lexical priors in regions where humans themselves show weaker agreement about color categories.

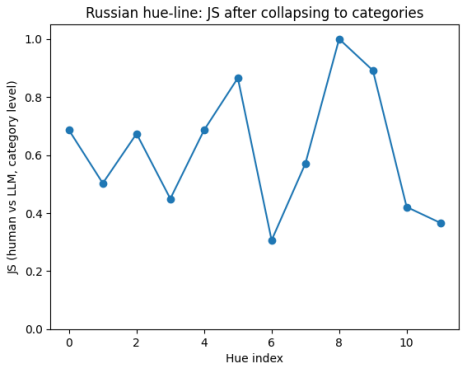

Figure 2: JS divergence on Russian hue-line categories (Exp. C).

Lang	Bins	JS_s	JS_h	IV_s	IV_h	VR_h
ZH	150	0.762	0.217	0.180	0.692	0.472
DE	48	0.460	0.365	0.406	0.524	0.307
KO	150	0.563	0.474	0.338	0.409	0.190
PL	36	0.602	0.589	0.364	0.390	0.661
RU	80	0.726	0.641	0.231	0.337	0.795
FR	27	0.739	0.744	0.186	0.227	0.419
EN	150	0.770	0.821	0.109	0.117	0.043
ES	92	0.869	0.876	0.084	0.090	0.549
PT	50	0.925	0.965	0.060	0.025	0.795

Table 5: Experiment D: Per-language distributional agreement on full CIELAB grids. JS_s/JS_h = surface/head Jensen–Shannon divergence. IV_s/IV_h = in-vocabulary human mass. VR_h = head-level vocabulary ratio.

4.4 Experiment D: Full CIELAB Grid Across Nine Languages

This experiment evaluates distributional agreement between GPT-4o and human naming over the full binned CIELAB grid. As described in Section 3.5, we sample GPT-4o 3–5 times per bin; JS is interpreted comparatively. Figure 3 shows example RGB tiles presented to GPT-4o in Experiment D, sorted by approximate LAB chroma.

Table 5 reports mean surface and head-level JS divergence and in-vocabulary mass by language. VR_h reports the mean head-level vocab_ratio across retained bins per language. In-vocabulary mass denotes the fraction of human probability mass under the normalized human distribution $p_{\text{human}}(t)$ that is assigned to labels appearing in the model output. Values are averaged across retained bins per language.

To test whether divergence is associated with reduced chroma, we computed Spearman correlations between CIELAB chroma $C^* = \sqrt{a^{*2} + b^{*2}}$ and head-level JS divergence. Several languages, including English ($\rho = -0.22$, $p = .008$), Chinese ($\rho = -0.19$, $p = .017$), and Spanish ($\rho = -0.24$, $p = .019$), showed weak negative correlations, indicating somewhat higher divergence in lower-chroma regions. However, this relationship was

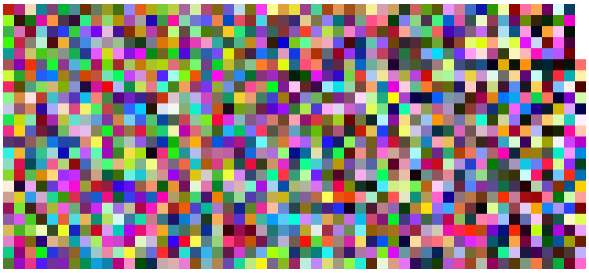

Figure 3: Experiment D: Example full-grid stimuli drawn from the retained CIELAB bins. Each tile is one bin shown to GPT-4o.

inconsistent across languages and absent in several cases, suggesting that chroma alone does not fully explain model–human divergence.

By contrast, the proportion of human responses assigned to the dominant head category was strongly negatively correlated with JS divergence across all languages (e.g., English $\rho = -0.75$, Russian $\rho = -0.54$, German $\rho = -0.90$). Bins with strong human category agreement therefore tended to show much closer model–human agreement.

4.4.1 Cross-Linguistic Summary

Table 5 summarizes Experiment D across nine languages. We focus on the six high-coverage (data-rich) languages (ZH, EN, KO, DE, RU, ES); FR/PL/PT are treated as exploratory because their coverage and/or annotation density after filtering is less stable for the present analyses. For a bin with in-vocabulary mass m, the fraction of human mass on labels never produced by the model is $1-m$. Many languages show low in-vocabulary mass (often below 0.2), meaning most human probability mass lies on labels absent from the model's sampled outputs.

4.4.2 Error Types and Bin Quality

We classify bins into *good*, *mid*, or *bad* using JS_{head} thresholds. Table 6 shows per-language counts. Chinese displays the lowest divergence with many "good" bins, while Korean has the largest number of bad bins, followed by Spanish and English.

Lang	#Bins	Good	Mid	Bad
Chinese	150	107	28	15
English	150	71	55	24
French	27	9	13	5
German	48	24	15	9
Korean	150	42	57	51
Polish	36	11	14	11
Portuguese	50	21	21	8
Russian	80	25	37	18
Spanish	92	33	32	27

Table 6: Experiment D: counts of bins labeled *good*, *mid*, or *bad* (good ≤ 0.3, bad ≥ 0.7).

Error patterns are consistent across languages:

- **High-chroma bins** (reds, greens, blues) are

usually aligned with human naming.

- **Low-chroma bins** account for many "bad" cases and show fallback to gray or beige labels.
- **Mid bins** reflect small shifts across neighboring human categories.

4.4.3 Confusion Patterns

Typical errors include overuse of desaturated labels (e.g., *gray*, *серый*, 회색), collapsing multiple human blues (e.g., *sky blue*, *turquoise*, *голубой*) into a single head, and pink/purple spillover in neighboring chromatic regions. These behaviors match the vocabulary-compression statistics in Tables 5 and 6 and recur across languages.

4.5 Control Experiments: Text, Saturation, and Temperature

The control experiments address three possible confounds: purely linguistic priors, weak chroma sensitivity, and sampling temperature.

Text-only baseline. Removing the image and prompting GPT-4o in text-only mode produces distributions that are almost maximally different from human naming on the hue-line tiles, as shown in Table 7. For English, JS divergence between human and text-only model distributions lies in a narrow and very high range ($0.971 \leq \text{JS} \leq 0.982$, mean 0.977). For Russian, divergence is slightly lower but still very high ($0.833 \leq \text{JS} \leq 0.949$, mean 0.902). The intrinsic color-name prior in both languages is therefore not human-like. Any partial agreement with human naming in the image-based experiments cannot be attributed to linguistic priors alone.

Saturation ratings. Table 8 shows the results of the saturation sanity check. GPT-4o shows only weak separation between high- and low-chroma patches. In English, mean ratings for both groups are essentially identical (both near 2.0 on a 1–5 scale). In Russian, high-chroma tiles receive mean ratings around 2.0 and low-chroma tiles around 1.7, with substantial variability and overlapping ranges. Across both languages, GPT-4o rarely assigns ratings near the extremes of the scale and does not treat high-chroma stimuli as clearly more saturated than low-chroma ones. These results are consistent with weak chroma sensitivity for uniform patches, though rendering and preprocessing effects cannot be ruled out.

Temperature sweep. The low-chroma temperature sweep shows that the gray-collapse effect is robust to sampling settings (Table 9). For English, mean JS divergence on the ten selected low-chroma tiles remains extremely high at all temperatures (mean JS 0.985 at $T = 0.1$, 0.979 at $T = 0.3$, 0.977 at $T = 0.7$, and 0.978 at $T = 1.0$). Russian shows slightly lower but still very high divergence (mean JS 0.956, 0.952, 0.946, and 0.928 for the same temperatures). Temperature adjustments do not rescue model performance in low-chroma regions. GPT-4o continues to collapse diverse human chromatic responses into a small set of grayish labels.

5 Discussion

Across languages and stimulus types, GPT-4o shows a stable pattern. The model has strong lexical knowledge of high-chroma colors and reproduces several language-specific distinctions at the category-head level (for example, Russian *голубой–синий* and German *blau–grün*) when category boundaries are unambiguous, consistent with heavy text-based training.

The contrast between the synthetic hue wheel (Experiment A) and the human-derived CIELAB grids (Experiment D) is partly associated with low-chroma regions, although the relationship between chroma and divergence is weak and inconsistent across languages. Experiment A contains only vivid, high-saturation hues with strong visual signals, whereas the full-grid evaluation includes many low- and mid-chroma regions where cues are weak. Under these conditions, GPT-4o frequently produces desaturated labels such as "gray" or "beige," even when human naming remains chromatic. Because the same RGB rendering pipeline is used across experiments (with HSV or CIELAB as color sources), these discrepancies are most plausibly related to weak or insufficiently exploited chroma cues rather than implementation differences.

Across low-chroma settings (Experiments C and D), the model displays the same qualitative behavior: it replaces diverse human labels with a small set of high-frequency, desaturated terms.

The control experiments sharpen this interpretation. The text-only baseline shows that GPT-4o's purely linguistic color-name prior is highly divergent from human naming on low-chroma tiles, while the saturation sanity check indicates weak separation between high- and low-chroma patches, and the temperature sweep shows that fallback to

Lang	Mean JS	Std	Min JS	Max JS
English	0.977	0.003	0.971	0.982
Russian	0.902	0.030	0.833	0.949

Table 7: Text-only baseline on hue-line tiles.

Lang	Chroma	Mean	Std
EN	high	2.00	0.00
EN	low	2.07	0.12
RU	high	2.00	1.00
RU	low	1.67	0.58

Table 8: Saturation rating sanity check.

Lang	T	Mean JS	Std	Min–Max
EN	0.1	0.985	0.013	0.971–1.000
EN	0.3	0.979	0.009	0.966–0.999
EN	0.7	0.977	0.009	0.967–0.999
EN	1.0	0.978	0.009	0.968–0.995
RU	0.1	0.956	0.038	0.884–1.000
RU	0.3	0.952	0.037	0.874–1.000
RU	0.7	0.946	0.031	0.880–1.000
RU	1.0	0.928	0.020	0.890–0.960

Table 9: Low-chroma temperature sweep.

desaturated labels persists regardless of sampling settings. These results suggest that GPT-4o appears to rely more heavily on lexical priors in regions where human category structure is weaker or less concentrated. GPT-4o is dependable for vivid, high-chroma colors but inconsistent for low- and mid-chroma stimuli. The text-only baseline and control experiments indicate that the model shows limited behavioral sensitivity to fine chroma distinctions in this setup, and applications requiring subtle color judgments should therefore treat current multimodal LLMs as unreliable in low-chroma regions without explicit calibration to human data.

6 Conclusion

GPT-4o shows lexical knowledge of color terms and reproduces language-specific distinctions under constrained prompting, but across nine languages and hundreds of CIELAB bins, it does not reliably approximate human color-naming distributions. The contrast between strong lexical availability and weaker distributional agreement is most pronounced in low-chroma and boundary regions. These results suggest that current multimodal language models remain less reliable for subtle color judgments and may benefit from training regimes that more tightly couple language with perceptual supervision.

Limitations

Coverage across languages is limited by the availability of human naming datasets; eight languages include $\geq$ 30 bins for robust comparison, and results for French (27 bins) should be treated as exploratory.

Image rendering is standardized but not calibrated to display hardware or to the exact conditions of the original human experiments; absolute color appearance may therefore differ between our setup and the one used to collect the UW data.

GPT-4o sampling cost constrained us to sparse sampling in the human-derived evaluations: Experiment D uses 3–5 draws per bin, while the hue-line and control experiments use larger but still limited sample counts.

Our conclusions about human distributional agreement are therefore interpretive rather than causal: the present design cannot fully rule out rendering artifacts, temperature effects, or prompt biases as contributing factors.

Finally, this work does not investigate the internal representations or generation dynamics that contribute to vocabulary collapse; future work should examine model internals more directly.

Because retained bins differ across languages after filtering, chroma ranges are not identical across language-specific analyses.

Ethical Considerations

This study analyzes a pretrained multimodal model using synthetic stimuli and existing, anonymized human color-naming datasets. No new human subjects were involved and no personal or sensitive data were collected. The work poses minimal direct ethical risk and is intended to inform evaluation and model design rather than normative claims about perception or language.

Data and Code Availability. Code for stimulus generation, prompting, normalization, evaluation, and figure generation is available at https://github.com/bondfeld/multilingual-color-naming-gpt4o. The human color-naming data are derived from the publicly available UW Color Names corpus (Kim et al., 2019).

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds,

Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, and 8 others. 2022. Flamingo: a visual language model for few-shot learning. *Preprint*, arXiv:2204.14198.

Brent Berlin and Paul Kay. 1969. *Basic Color Terms: Their Universality and Evolution*. University of California Press.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. 2020. Experience grounds language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8718–8735, Online. Association for Computational Linguistics.

Rémi Cadène, Corentin Dancette, Hédi Ben-Younes, Matthieu Cord, and Devi Parikh. 2019. Rubi: Reducing unimodal biases in visual question answering. *CoRR*, abs/1906.10169.

Jingyi Chen, Zhimeng Guo, Jiyun Chun, Pichao Wang, Andrew Perrault, and Micha Elsner. 2026. Do audio LLMs really LISTEN, or just transcribe? measuring lexical vs. acoustic emotion cues reliance. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5848–5877, Rabat, Morocco. Association for Computational Linguistics.

Stevan Harnad. 1990. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1–3):335–346.

Kimberly Jameson and Roy D'Andrade. 1997. It's not really red, green, yellow, blue: an inquiry into perceptual color space. In C. L. Hardin and Luisa Maffi, editors, *Color Categories in Thought and Language*, pages 295–319. Cambridge University Press.

Paul Kay and Terry Regier. 2003. Resolving the question of color naming universals. *Proceedings of the National Academy of Sciences*, 100(15):9085–9089.

Younghoon Kim, Kyle Thayer, Gabriella Silva Gorsky, and Jeffrey Heer. 2019. Color names across languages: Salient colors and term translation in multilingual color naming models. In *Proceedings of EUROVIS 2019 –Short Papers*. The Eurographics Association.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*.

Yijun Liang, Ming Li, Chenrui Fan, Ziyue Li, Dang Nguyen, Kwesi Cobbina, Shweta Bhardwaj, Jiuhai Chen, Fuxiao Liu, and Tianyi Zhou. 2025. Colorbench: Can vlms see and understand the colorful world? a comprehensive benchmark for color perception, reasoning, and robustness. *arXiv preprint arXiv:2504.10514*.

Will Monroe, Robert X. D. Hawkins, Noah D. Goodman, and Christopher Potts. 2017. Colors in context: A pragmatic neural model for grounded language understanding. *Transactions of the Association for Computational Linguistics*, 5:325–338.

OpenAI. 2023. Gpt-4 technical report. Technical report, OpenAI. ArXiv:2303.08774.

OpenAI. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*. Accessed: 2026-05-20.

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

Roger N. Shepard. 1992. The perceptual organization of colors: An adaptation to regularities of the terrestrial world? In Jerome H. Barkow, Leda Cosmides, and John Tooby, editors, *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*, pages 495–532. Oxford University Press, New York.

Guillaume Thierry, Panos Athanasopoulos, Alison Wiggett, Benjamin Dering, and Jan-Rouke Kuipers. 2009. Unconscious effects of language-specific terminology on preattentive color perception. *Proceedings of the National Academy of Sciences*, 106(11):4567–4570.

Jonathan Winawer, Nathan Witthoft, Michael C. Frank, Lisa Wu, Alex R. Wade, and Lera Boroditsky. 2007. Russian blues reveal effects of language on color discrimination. *Proceedings of the National Academy of Sciences*, 104(19):7780–7785.

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

A Appendix: Prompts

A.1 Experiment A: Open-vocabulary prompt

English You see a solid-colored square. Give a natural, descriptive English color name. Avoid using only basic color words like red, orange, yellow, green, blue, or purple. Use a more specific or nuanced color term, such as 'scarlet', 'deep red', 'lime green', 'teal', 'aquamarine', 'magenta', or a similar descriptive phrase. Answer with one word or a very short phrase.

Russian Вы видите квадрат, залитый одним цветом. Дайте естественное русское название этого цвета. Не ограничивайтесь только базовыми цветами (красный,

оранжевый, жёлтый, зелёный, синий, голубой, фиолетовый). Используйте более точное или оттеночное название, например ’алый’, ’бордовый’, ’бирюзовый’, ’лазурный’, ’лиловый’ и т.п. Ответьте одним словом или очень короткой фразой.

A.2 Experiment B: Fixed categories

English You see a solid-colored square. Name its BASIC color category in English. Use exactly ONE word from this list: red, orange, yellow, green, blue, purple, pink, brown, gray, black, white. Answer with just that single word.

Russian Вы видите одноцветный квадрат. Назовите ЕГО ОСНОВНУЮ КАТЕГОРИЮ ЦВЕТА по-русски.Используйте ровно ОДНО слово из этого списка: красный, оранжевый, жёлтый, зелёный, синий, голубой, фиолетовый, розовый, коричневый, серый, чёрный, белый. Ответьте только одним словом.

A.3 Experiment C: Hue-line Divergence

English You see a solid-colored square. Give a natural, descriptive English color name. Avoid using only basic color words like red, orange, yellow, green, blue, or purple. Use a more specific or nuanced color term, such as ’scarlet’, ’deep red’, ’lime green’, ’teal’, ’aquamarine’, ’magenta’, or a similar descriptive phrase. Answer with one word or a very short phrase.

Russian Вы видите квадрат, залитый одним цветом. Дайте естественное русское название этого цвета. Не ограничивайтесь только базовыми цветами (красный, оранжевый, жёлтый, зелёный, синий, голубой, фиолетовый). Используйте более точное или оттеночное название, например «алый», «бордовый», «бирюзовый», «лазурный», «лиловый» и т.п. Ответьте одним словом или очень короткой фразой.

A.4 Experiment D: Full CIELAB Grid

Experiment D uses the same open-vocabulary prompt family as Experiments A and C, instantiated in nine languages (English, Russian, Chinese, Korean, German, French, Spanish, Polish, Portuguese). For each CIELAB bin, GPT-4o sees a solid color patch and is asked for a single natural color name in the target language, with no language mixing or explanation.

Lang	#Subjects
ZH (Chinese)	5686
EN (English)	145709
FR (French)	2968
DE (German)	3721
KO (Korean)	13507
PL (Polish)	1085
PT (Portuguese)	1661
RU (Russian)	1682
ES (Spanish)	4124

Table 11: Unique human participants per language in the UW color-naming dataset.

Generic template (all languages) You see a solid-colored square. Give a natural, descriptive [target language] color name. Answer with one word or a very short phrase in [target language] only.

English (used in A, C, and D) You see a solid-colored square. Give a natural, descriptive English color name. Avoid using only basic color words like red, orange, yellow, green, blue, or purple. Use a more specific or nuanced color term, such as “scarlet”, “deep red”, “lime green”, “teal”, “aquamarine”, “magenta”, or a similar descriptive phrase. Answer with one word or a very short phrase.

Russian (used in A, C, and D) Вы видите квадрат, залитый одним цветом. Дайте естественное русское название этого цвета. Не ограничивайтесь только базовыми цветами (красный, оранжевый, жёлтый, зелёный, синий, голубой, фиолетовый). Используйте более точное или оттеночное название, например «алый», «бордовый», «бирюзовый», «лазурный», «лиловый» и т.п. Ответьте одним словом или очень короткой фразой.

Additional Languages For Chinese, Korean, German, French, Spanish, Polish, and Portuguese, we use direct translations of the generic template, with the language name and example color terms adapted to the target language.

B Appendix: Additional Figures

Language	Full form	Head
EN	light blue	blue
EN	dark green	green
RU	ярко-синий	синий
RU	светло-зелёный	зелёный
ES	verde claro	verde

Table 10: Examples of head–modifier normalization.

B.1 Experiment C: Hue-Line Divergence

B.2 Experiment D: Full CIELAB Grid

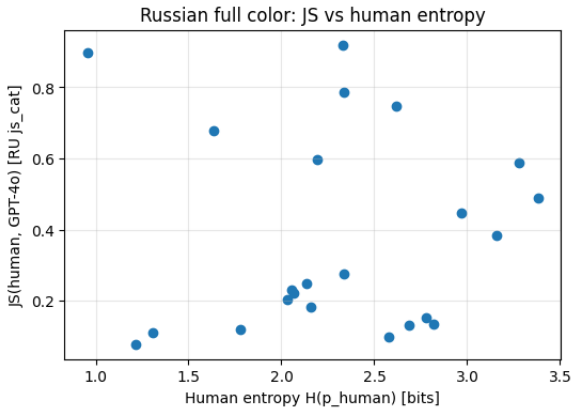

Figure 4: Russian full-grid: human entropy vs JS divergence.

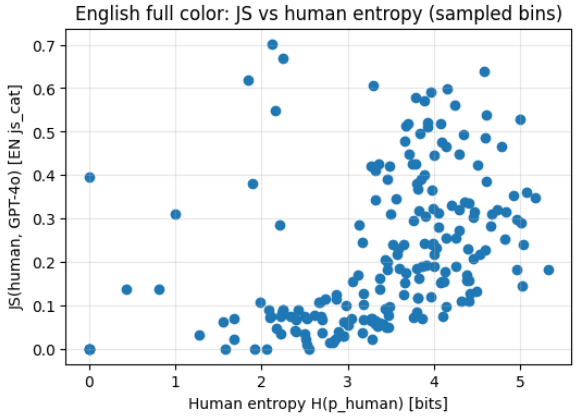

Figure 5: English full-grid: human entropy vs JS divergence.

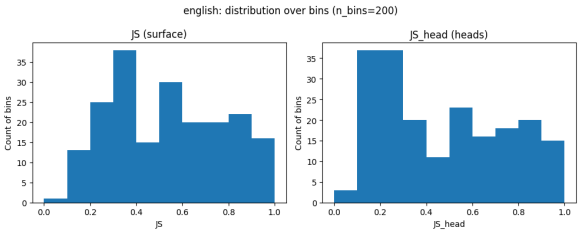

Figure 6: JS_h distributions for English (Exp. D).

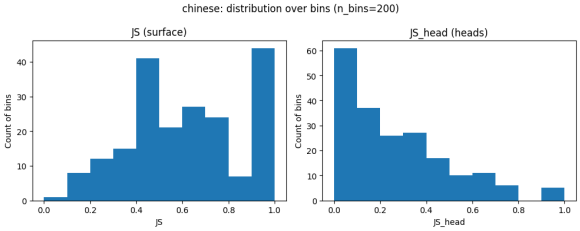

Figure 7: JS_h distributions for Chinese (Exp. D).

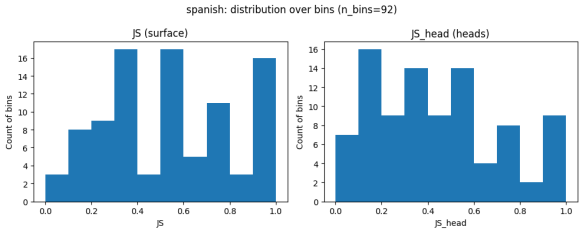

Figure 8: JS_h distributions for Spanish (Exp. D).

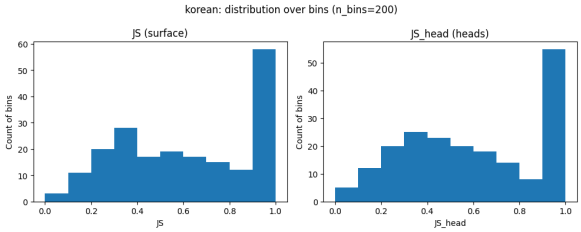

Figure 9: JS_h distributions for Korean (Exp. D).

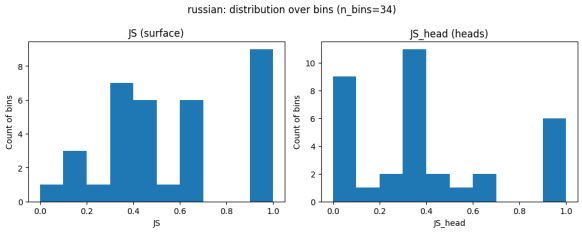

Figure 10: JS_h distributions for Russian (Exp. D).