The Shades of Meaning: Investigating LLMs' Cross-lingual Representation of Grounded Structures

Stanford CS224N Custom Project

Pinlin [Calvin] Xu
Department of Computer Science
Stanford University
pinlinxu@stanford.edu

Garbo Chung
Department of Computer Science
Stanford University
garbo22@stanford.edu

Abstract

Are Large Language Models stochastic parrots that recite the most probable answer, or do they learn representations of grounded structures in the real world to answer your questions well? In this study, we investigate the claim that LLMs learn and linearly encode models of the world with a case study in colors. Guided by a series of falsifiable hypotheses, we perform a series of baselines and controlled experiments to discover how the quality of representations varies with language, model, context, and fine-tuning.

We find that English contextual embeddings produced by LLMs do not meaningfully outperform the glove-wiki-gigaword-300 baseline for smaller models (7 to 14 billion parameter), and that LLMs consistently possess better representations for English color terms, even for a model trained predominantly on Japanese. Nevertheless, alignment of representations with the underlying color space improves with both finetuning and parameter size, and multilingual LLMs could be valuable tools for analyzing crosslingual encoding of perceptions. To aid this, we introduce a color-mapping experiment that visualizes languages' color representation, and offer our analyses and interpretations from linguistic anthropology perspectives.

1 Key Information

- Mentors: Sonia Hangjie Chu (TA)
- External collaborators: No
- External mentors: Isabel Papadimitriou (Stanford NLP Group), Karen Livescu
- Team Contributions: see below
- Late days: 1 for milestone; at most 3 for report; each member supplies 4 days

nullus enim color est omnino materiai
corporibus, neque par rebus neque denique dispar
in quae corpora si nullus tibi forte videtur
posse animi iniectus fieri, procul avius erras

for the bodies of matter have no colour at all
neither like things nor again unlike them
but if by chance you think that the mind cannot
project itself into these bodies, you wander far astray

Lucretius, *De Rerum Natura* 2.737. Trans. Bailey (1921)

2 Introduction

Large Language Models (LLMs) demonstrate notable capabilities in natural language understanding. However, the extent to which they grasp the meaning of the language they process remains under heavy scrutiny (i.e., the "stochastic parrot" argument). Critics often argue that LLMs lack both symbolic structure and grounding, and on the latter point that the language modeling task is inadequate for learning language representation as it relates to the real, physical world (Pavlick, 2023), (Bender and Koller, 2020).

While this debate is deeply philosophical and depends on the type of semantics being discussed[1], recent research exploring the connections between LLMs' hidden states and grounded structures—such as color, directions, time, or geographic positions—has yielded compelling findings. Although no formal proof yet confirms that the representations learned by LLMs are isomorphic to these structures, evidence of robust, often

[1]In the final lecture of CS 224N Spr 2024 titled "NLP, Linguistics, and Philosophy", Professor Manning seems to reject both points, stating that the brain is a neural NLP system evidently capable of processing symbolic systems, and though the "real world" and classic grounded meaning is privileged, it is not a superset of meaning derived from language context, per the "shehnai" example.

linear mappings and strong correlations have led to claims such as that "LLMs learn linear representations [...] of the real world and possess basic ingredients of a world model" (Gurnee and Tegmark, 2024). This field of investigation is emerging and rapidly evolving; the studies within it, noted for their novelty, have also faced criticism for making broad generalizations across different models and data (Reviewers, 2024), lacking baseline comparisons (Ruder, 2023), and conducting analyses in English only.

This project aims to address the aforementioned concerns. By setting up various baselines and a series of falsifiable hypotheses, we aim to impartially analyze the results from controlled, multilingual experiments relate the results from prior studies to informative baselines. Indeed somewhat disappointingly, our results show that LLM color representations (namely contextual embeddings) do not meaningfully outperform pretrained GloVe embeddings in correlation with the underlying physics-based color space, except for the largest and latest Llama3-70B-Instruct model. Nevertheless, they provide the means to easily acquire cross-lingual contextual representations that enable novel analysis of culture and language usage.

Team contributions: Equal. Pinlin [Calvin] Xu mainly worked on compiling data, prototyping experiments, finetuning Llama3, and visualizing color mapping; Garbo Chung mainly worked on extracting contextual color representations from models, and performing grounding evaluation experiments.

3 Related Work

3.1 Large Language Models' Grounded Representations

In the pioneering work Abdou et al. (2021), the authors applied BERT (Devlin et al., 2019) to analyze how 51 American English speakers named Munsell color chips using monolexemic color terms. Using both Representation Similarity Analysis (RSA) (Kriegeskorte et al., 2008) and the fitness of LASSO-regularized linear regression, they assessed the alignment between 1024-dimensional contextual embeddings for color terms and 3D CIELAB color coordinates.

Subsequent research have expanded the evaluation to more diverse structures, such as spatial and temporal coordinates (Gurnee and Tegmark, 2024), directions (Patel and Pavlick, 2022), or more abstractly object properties (Forbes et al., 2019) and falsehood (Marks and Tegmark, 2023). They have generally focused on decoder-only autoregressive language models, likely motivated by both the increased prominence of LLMs and the usage of specialized research models like Pythia (Biderman et al., 2023) to explore additional model sizes and training dynamics.

Gurnee and Tegmark (2024) provides the most comprehensive testing to date, probing on various combinations of model sizes and hidden layers. Significant correlation, increasing with layer depth, is reported, and deemed linear as nonlinear mappings like MLPs show minimal improvement. Some reviewers, however, find their results not particularly surprising: pretrained word embeddings such as Word2Vec (Mikolov et al., 2013b) and GloVe (Pennington et al., 2014) have long demonstrated linear algebraic structures of word meanings; and still earlier Louwerse and Zwaan (2009) reported similarly extracting longitude and latitude coordinates from Latent Semantic Analysis (LSA) embeddings, yet no traditional baselines were not compared against (Ruder, 2023).

Besides such probing studies, research rooted in cognitive science has introduced "behavioral" studies that recover representations from LLMs'output tokens: Kawakita et al. (2023) measures response pattern similarities and Zhu et al. (2024) applies established sampling techniques such as Markov Chain Monte Carlo with People (MCMCP) on GPT-4. Nevertheless the prospect of identifying structures directly from the hidden state remains compelling, with emerging research like (Alper et al., 2023) focusing on the effects of vision and language pretraining; notably Merullo et al. (2023) has shown linear mappings between representations learned by vision-only and text-only models.

3.2 Crosslingual Embeddings and Alignment

Aligned cross-lingual representations in a joint embedding space are crucial for NLP tasks like translation and retrieval, and for language transfer learning (Ruder et al., 2019). As pioneered by Mikolov et al. (2013a), assuming embedding spaces are approximately isomorphic across languages, term translation can be accomplished by mapping monolingual embeddings to a shared space using orthogonal transformations and finding the nearest neighbor in the target language via cosine similarity (Søgaard et al., 2018) (Kementchedjhieva et al., 2018). While traditional approaches required a cross-lingual reference (e.g., parallel corpus) to "anchor" alignment, increasingly sophisticated methods such as Conneau et al. (2018) moved towards increasingly unsupervised processes that used or jointly trained embeddings from Skip-Gram Negative Sampling (SGNS) before the advent of LLMs. Contextual embeddings, such as those from ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019), were found to consistently outperform previous state-of-the-art methods post alignment (Schuster et al., 2019; Wang et al., 2019). It is further hoped that LLMs, trained on unprecedented amounts of multilingual data, will enhance the analysis of cross-cultural encoding of visual, spatial, and temporal perceptions (Ruder, 2023).

4 Approach

Consider the general architecture of transformer-based language models: input tokens are converted to vectors in the initial embedding layer, combined with positional embeddings, and transformed over successive attention blocks until they are finally converted back to tokens (Radford and Narasimhan, 2018). We consider the contextual embeddings generated by the language model to be the final vectors prior to the output layer.

Our main task is investigating the alignment between high-dimensional embeddings of color terms and ground truth color coordinates in a specific color space (e.g., RGB, CIELABS, CMYK, etc.) Given a concrete scenario where:

- $\mathbf{X} \in \mathbb{R}^{N \times p}$ as the input word embeddings, where $p = 4096$ for most models.
- $\mathbf{y} \in \mathbb{N}^{N \times q}$ as the (clamped) output RGB color values, where $q = 3$ for most color spaces.

We introduce three evaluation metrics. First, we perform regression analysis via probing, introduced by Alain and Bengio (2016) and a consistent metric in prior literature since Abdou et al. (2021). We train a linear mapping from the embedding space in $\mathbb{E}^{4096}$ to the RGB color space in $\mathbb{E}^3$ with the following Elastic Net regularization where

$$\mathcal{L}(\mathbf{W}, \mathbf{b}) = \frac{1}{N} \sum_{i=1}^{N} \|\mathbf{y}^{(i)} - \widehat{\mathbf{y}}^{(i)}\|_2^2 + \lambda_1 \|\mathbf{W}\|_1 + \lambda_2 \|\mathbf{W}\|_2^2 \tag{1}$$

- $N = 2054 \cdot 0.8$ is the number of training samples from the main English-Chinese bilingual dataset compiled.
- λ_1 and λ_2 are regularization parameters for L1 and L2 regularization, obtained via grid-search and leave-out cross-validation with at most 3000 iterations. $\mathbf{W} \in \mathbb{R}^{3 \times 4096}$, $\mathbf{b} \in \mathbb{R}^3$, $\widehat{\mathbf{y}} \in \mathbb{R}^3$

Regression fitness, namely the coefficient of determination $R^2 = 1 - \frac{\sum_{i=1}^{N} \|\mathbf{y}^{(i)} - \widehat{\mathbf{y}}^{(i)}\|_2^2}{\sum_{i=1}^{N} \|\mathbf{y}^{(i)} - \bar{\mathbf{y}}\|_2^2}$, is reported on alignment. Additional metrics are density ($\frac{|\{w_i \neq 0\}|}{|\mathbf{W}|}$) and mean regression error in terms of normalized Euclidean distance defined below

$$d_{\text{norm}}(\mathbf{c}_{\text{actual}}, \mathbf{c}_{\text{predicted}}) = \frac{\|\mathbf{c}_{\text{actual}} - \mathbf{c}_{\text{predicted}}\|_2}{\|\mathbf{c}_{\text{white}} - \mathbf{c}_{\text{black}}\|_2} \tag{2}$$

where $\|\mathbf{c}_{\text{white}} - \mathbf{c}_{\text{black}}\|_2 = \sqrt{255^2 \cdot 3} \approx 441.673$ to help interpret the relative difference.

Second, we conduct Canonical-Correlation Analysis (CCA) to identify linear combinations of $\mathbf{X}$ and $\mathbf{y}$ (specifically the canonical random variables $\mathbf{a}_k^T \mathbf{X}$ and $\mathbf{b}_k^T \mathbf{y}$, where $\mathbf{a}_k \in \mathbb{R}^p$ and $\mathbf{b}_k \in \mathbb{R}^q$ are coefficients known as the canonical vectors) that maximize the correlation $\rho = \text{corr}(\mathbf{a}_k^T \mathbf{X}, \mathbf{b}_k^T \mathbf{y})$ (Zhuang et al., 2020).

However, directly applying CCA presents issues: consider the row-space of the data matrix $\mathbf{Z} = [\mathbf{X}, \mathbf{y}]^T \in \mathbb{R}^{N \times (p+q)}$. This system becomes overdetermined if $p + q \geq N$, which holds due to large model hidden sizes and limited data, and we observed some canonical correlations are always 1 (Shokouhi, 2023). To address this, dimensionality reduction is applied, using the linear method of Principal Component Analysis (PCA) or the nonlinear method of Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2018), both commonly used for embedding visualization and analysis. Thus given

$$(\mathbf{X}_c, \mathbf{y}_c) = \text{CCA}(\mathbf{X}_{\text{reduced}} = [(\text{PCA or UMAP})(\mathbf{X}, q)], \mathbf{y}, q) \tag{3}$$

we report the canonical correlations $\rho_i = \text{corr}(\mathbf{X}_c[:, i], \mathbf{y}_c[:, i])$ denoting the Pearson correlation coefficient.

4.1 Color Mapping Visualization

Motivated by previous research in crosslingual embedding alignment, we introduce a pipeline to visualize differences in pairwise color representations, as shown in Figure 1. Given color term embeddings from a

LLM in two languages, we align them to a joint space using Procrustes analysis, involving solving for an orthogonal matrix that optimally maps the subset embeddings of Basic Color Terms, as identified by the World Color Survey, from one language to another (note that this system is also highly underdetermined) (Kay and Cook, 2014). The three basic colors black, white, and red are used in final results.

We hypothesize that if LLMs learn robust and accurate color representations across multiple languages, such representations after alignment should be similar, aside from random training variances and cultural nuances inherent to the languages and corpora used. While color values closely matching the ground truth can be recovered from the embeddings of either language, when we apply a regressor fitted to one language' s embeddings onto another, we observe significant differences. These differences alter the perception and mood of an image when the color mapping is applied, though their interpretation is difficult and subtle.

We do not reuse any existing codebase, other than ML libraries including HuggingFace Transformers (Wolf et al., 2020), Pytorch (Paszke et al., 2019), scikit-learn (Pedregosa et al., 2011), SciPy (Virtanen et al., 2020), gensim (Řehůřek and Sojka, 2010) to conduct original experiments.

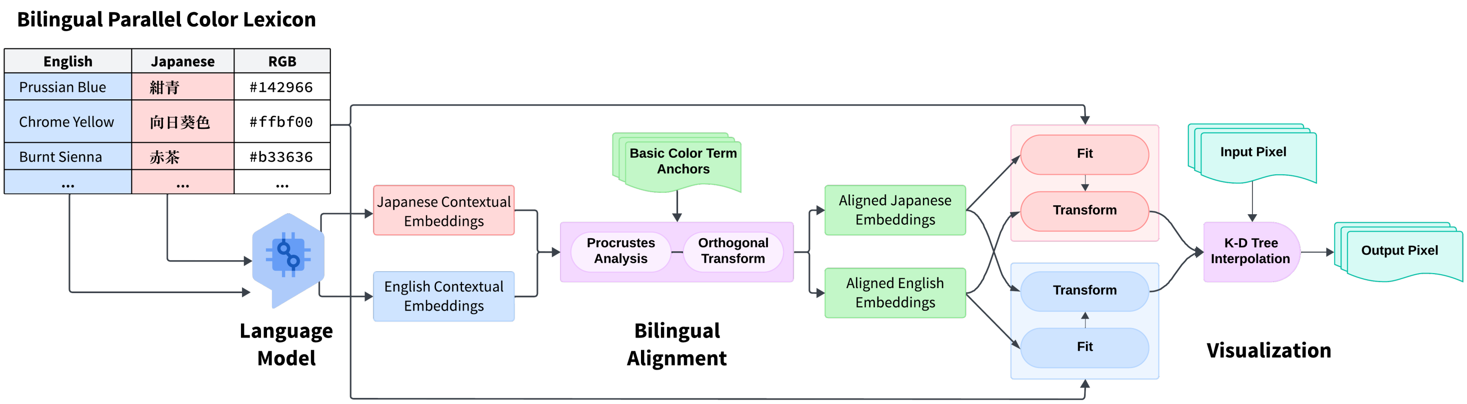

Figure 1: Color Mapping Visualization Pipeline

5 Experiments

5.1 Data

We have collected a total of 17566 named colors and their definition in some color space(s) from various sources. A valuable subset of them is bilingual: for example, Table 1 a color entry defined in the RAL Classic color standard (Europe, 2024). Additionally, Table 2 shows the rich natural language descriptions we collected for traditional Japanese and Chinese colors detailing their composition, cultural impression and usage, etc., that support our cultural analysis experiments.

English	Chinese	Hex	R	G	B	H°	S%	L%	L*%	a*%	b*%	C%	M%	Y%	K%
Saffron yellow	藏红花黄色	#F6A950	246	169	80	33.75°	84.21	62.75	75.183	20.633	55.581	0	29	67	6

Table 1: Example: RAL 1017 Saffron yellow; compiled from (Wikipedia contributors, 2023)

Kanji	Kana	Romaji	Hex	C	M	Y	K	Description
茜色	あかねいろ	Akane-iro	#B7282E	0	78	75	28	茜色（あかねいろ）とは、茜草の根で染めた暗い赤色のことです。夕暮れ時の空の形容などに良く用いられることで知られています...

Table 2: Example traditional Japanese color compiled from (koka); translation of the description such as: "'Akane-iro' refers to a dark red color dyed with the roots of *rubia cordifolia* (Indian madder). It is well known for being used to describe the color of the sky at dusk ... (continued)"

This data serve as a basis for our probing and embedding analysis experiments. Preprocessing is done to sanitize, deduplicate, remove nondescriptive names (such as "PMS 1485" from the Pantone Matching System) and fix errors from larger lists such as Krzywinski (2017). For most experiments, a high-quality English-Chinese bilingual subset with 2054 examples is used, and other usage will be reported per-case.

5.2 Experimental Details

The large language models we investigate are Mistral-7B-Instruct-v0.2 (Team), Llama-3-8B-Instruct (4bit quantized version from unsloth (2024)), Llama-3-70B-Instruct (AI@Meta, 2024), and Fugaku-LLM-13b-instruct that is specialized in Japanese (Fujitsu, 2024). Their embedding dimensions are 4096, 4096, 8192, and 5184, respectively, with vocabulary sizes of respectively 32000, 128256, 128256, 51200.

Parameter-Efficient Fine Tuning (PEFT) is performed on Llama-3-8B-Instruct to evaluate changes in representation, where pretrained weight matrices are updated using Low-Rank Approximation (LoRA) to reduce the number of weight updates (Hu et al., 2021). During full finetuning, the model is initialized to pre-trained weights Φ_0 and updated to $\Phi_0 + \Delta\Phi$ by maximizing some conditional language modeling objective

$$\max_{\Phi} \sum_{(x,y)\in\mathcal{Z}} \sum_{t=1}^{|y|} \log P_{\Phi}(y_t|x, y_{<t}) \tag{4}$$

Instead of optimizing for $\Delta\Phi$ directly, whose dimensionality is equal to $|\Phi_0|$, we compute a task-specific $\Delta\Phi(\Theta)$ parameterized by a smaller Θ. As $|\Theta| \ll |\Phi_0|$ and $\Phi = \Phi_0 + \Delta\Phi(\Theta)$, the objective changes to

$$\max_{\Theta} \sum_{(x,y)\in\mathcal{Z}} \sum_{t=1}^{|y|} \log P_{\Phi_0+\Delta\Phi(\Theta)}(y_t|x, y_{<t}) \tag{5}$$

Consider $\Delta W \subseteq \Delta\Phi$ is represented as a low-rank decomposition $\Delta W = BA$, where $B \in \mathbb{R}^{d\times r}$, $A \in \mathbb{R}^{r\times k}$ are trained on and $r \ll \min(d, k)$. The update to the some pretrained matrix $W_0 \in \mathbb{R}^{d\times k} \subseteq \Phi_0$ is then $W_0 + \frac{\alpha}{r}\Delta W = W_0 + \frac{\alpha}{r}BA$, which include `q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, `down_proj` in our setup. We set the hyperparameters as `rank` = `alpha` = 16, no dropout, a learning rate of 2×10^{-4}, and train for 3 epochs on 449 examples (training loss in Figure 6).

Experiments were conducted with compute available to us, including Tesla T4, TPU v2 instances via Google Colab, and an A100 (40GB) GPU instance from `modal.com`. The temperature for all embedding experiments was set to e^{-9}, using controlled contexts similar to the one described in 5.2.1, with mean-pooling applied to the final hidden layer. Some experiments were cut short due to Colab disconnects.

5.2.1 Evaluation Method & Retrieval Baseline Evaluation

Evaluation metrics used are discussed in depth in Section 4 Approach. Before investigating whether LLMs encode representations of grounded structures, we first verify if they act effectively as stochastic parrots, absent in previous studies. We perform a simple retrieval baseline for a small set of basic, well-defined colors in the HTML, CSS, and VGA standards ((Raggett et al., 1999)) with a structured prompt: `[INST] What is the RGB value of the following web color [whose name is in [Chinese, Japanese]]: [NAME OF COLOR]? [/INST]`

5.3 Results

5.3.1 Retrieval Baseline Results

We find that while Mistral-7B-Instruct-v0.2 performs almost flawlessly on this task when being asked in English, it appears quite undertrained in non-English languages, such as not knowing "fuchsia" in Chinese and responding with a color identical to purple. However, certain cultural factors are also evident. For example, the Japanese response to the color blue is a mixture of blue-green, due to a well-documented phenomenon that also leads to Japan still having some blue traffic lights (Iijima et al., 1982).

English	Chinese	Japanese	Definition	ref	en	zh	ja	en-d_{norm}	zh-d_{norm}	ja-d_{norm}
black	黑色	黒	`#000000`					0.0	0.0	0.0
silver	银色	銀	`#c0c0c0`					0.0	0.0	0.0
maroon	栗色	栗色	`#800000`					0.1820	0.1865	0.1770
red	红色	赤	`#ff0000`					0.0	0.0	0.0
navy	藏青色	ネイビー	`#000080`					0.0	0.645	0.0
blue	蓝色	青	`#0000ff`					0.0	0.0	0.4083
purple	紫色	紫	`#800080`					0.0	0.0	0.0
fuchsia	品红色	フクシャ	`#ff00ff`					0.0	0.5810	0.0
green	绿色	緑	`#008000`					0.0	0.0	0.0
lime	青柠色	ライム	`#00ff00`					0.0	0.4242	0.0
olive	橄榄绿	オリーブ	`#808000`					0.0	0.0977	0.1519
yellow	黄色	黄	`#ffff00`					0.0	0.0	0.0906
teal	蓝绿色	ティール	`#008080`					0.0	0.2728	0.0
cyan	青色	藍紫色	`#00ffff`					0.0	0.4201	0.5371
gray	灰色	灰色	`#808080`					0.0	0.0039	0.2510
white	白色	白	`#ffffff`					0.0	0.0	0.0

Figure 2: The 16 basic W3C web colors in different languages (Jaffer), and Mistral 7B's answers compared to definition.

5.3.2 Representation Probing Results and Analysis

As shown in Table 3, the regression analysis and canonical correlation results matched our expectations, showing that randomly permuted embeddings correlate poorly with their original colors, confirming that

Configuration	R^2 (out of sample)	$\bar{d}_{\text{norm}}$	Weight Density	$\vec{\rho}_{\text{PCA}}$
GloVe Embeddings	0.5952	**0.0915**	**0.13**	[**0.482**, **0.154**, **0.081**]
Label-Permuted Embeddings (×2)	-0.3735, -0.4480	0.1699, 0.1603	0.80, 0.89	[0.200, 0.072, 0.006], [0.198, 0.120, 0
Mistral English (Out of Context)	0.5192	0.1124	0.41	[0.250, 0.102, 0.036]
Mistral English (In Context)	**0.5969**	0.0980	0.21	[0.320, 0.069, 0.026]
Mistral English (In Context, Term Only)	0.5490	0.1037	0.73	[0.288, 0.143, 0.007]
Mistral Chinese (Out of Context)	0.3894	0.1128	0.69	[0.245, 0.135, 0.031]
Mistral Chinese (In Context)	0.4343	0.1044	0.49	[0.245, 0.135, 0.031]
Human (Author 1)	0.5630	0.1356	N/A	N/A
Human (Author 2)	-0.0299	0.2219	N/A	N/A

Table 3: Baseline Results

linear regression does not easily overfit and the coefficient of determination is recommended for assessing the alignment between embedding and color spaces. Despite low canonical correlations due to substantial dimensionality-reduction of either PCA or UMAP, relative comparisons can still be made.

As expected in context embeddings display better alignment. However, the pretrained glove-wiki-gigaword-300 embeddings perform favorably compared to even in-context English embeddings obtained from Mistral-7B-Instruct-v0.2 with minimal overfitting, prompting us to further investigate whether LLMs truly have significantly more robust and accurate representations of grounded structures compared to prior arts.

Additionally, both authors tested the regression baseline by selecting colors on an RGB wheel for 10 random bilingual entries (see Table A.1 in Appendix). Neither author matched the performance of the regression on GloVe embeddings, showing the inherent ambiguity of (some admittedly obscure) color names.

To assess whether LLMs exhibit a more grounded understanding of color in specific languages, we analyzed color embeddings from English, Chinese, and Japanese extracted from Llama-3-8b-Instruct-bnb-4bit. Continuing the trend from the baseline, we found that both the fitness of the regressor mapping embeddings to true RGB values and the canonical correlations were lower for Chinese and significantly lower for Japanese compared to English, suggesting Llama3 may have better understanding of colors in English.

Configuration	R^2	d_{norm}	Weight Density	$\vec{\rho}_{\text{PCA}}$	$\vec{\rho}_{\text{UMAP}}$
English Embeddings	**0.5544**	**0.0988**	0.92	**[0.488, 0.153, 0.127]**	**[0.346, 0.217, 0.012]**
Chinese Embeddings	0.4875	0.1054	0.92	[0.217, 0.181, 0.014]	[0.269, 0.130, 0.025]
Japanese Embeddings	0.2725	0.1325	**0.62**	[0.213, 0.143, 0.065]	[0.243, 0.089, 0.038]

Table 4: Llama3 Embedding Results

To assess whether a model trained predominantly in a non-English language would possess superior representations in that language, we analyzed the Japanese-trained Fugaku-LLM-13B-instruct, known for its training on a proprietary dataset of 380 billion tokens, 60% of which being Japanese (Fujitsu, 2024). Surprisingly, Japanese color embeddings from Fugaku showed similar or worse correlations and fitness compared to its English embeddings, and even compared to Japanese embeddings from Llama3. This suggests that the trend of LLMs having a better grounded understanding of English color terms is more persistent than we conjectured.

Configuration	R^2	d_{norm}	Weight Density	$\vec{\rho}_{\text{PCA}}$	$\vec{\rho}_{\text{UMAP}}$
English Embeddings	**0.5994**	**0.0930**	**0.68**	**[0.329, 0.149, 0.055]**	**[0.317, 0.173, 0.020]**
Japanese Embeddings	0.2701	0.1237	0.86	[0.158, 0.041, 0.002]	[0.229, 0.124, 0.060]

Table 5: Fugaku-LLM Embedding Results

To assess if finetuning improves LLMs'color representation alignment, we finetuned Llama3 on a dataset described in Table 2, consisting of Japanese cultural colors with extensive descriptions of their physical appearance and cultural usage. The model was tasked to predict their RGB definitions.

The finetuned model showed a slight improvement (bolded) in regressor fitness scores for both Japanese and Chinese embeddings compared to the original Llama3 model, with correlations also slightly higher or similar.

These modest gains, achieved over three epochs with about 400 colors, suggest that more extensive training could lead to significant improvements in understanding foreign cultural colors, as positive scaling with both training and parameter count has been reported by (Gurnee and Tegmark, 2024).

Configuration	R^2	d_{norm}	Weight Density	$\vec{\rho}_{\text{PCA}}$	$\vec{\rho}_{\text{UMAP}}$
English Embeddings	0.5544	**0.0980**	0.92	[**0.585**, 0.140, 0.101]	[**0.437**, **0.235**, **0.037**]
Chinese Embeddings	**0.5049**	**0.1053**	**0.84**	[**0.299**, 0.182, 0.001]	[0.196, 0.110, 0.001]
Japanese Embeddings	**0.2813**	**0.1301**	0.82	[**0.238**, **0.188**, 0.058]	[**0.332**, **0.260**, **0.041**]

Table 6: Finetuned Llama3 Embedding Results

To assess if representation quality scales with model size, we analyzed English embeddings from Llama-3-70B-Instruct. This model outperformed our finetuned 8B model and finally achieved a higher regressor fitness score than the glove-wiki-gigaword-300 baseline. Although a comprehensive series of experiments was not feasible due to constraints on time and computing resources, these results seem to support the scaling laws reported for Llama2 and Pythia models across various sizes Gurnee and Tegmark (2024).

Configuration	R^2	d_{norm}	Weight Density	$\vec{\rho}_{PCA}$	$\vec{\rho}_{UMAP}$
English Embeddings	**0.6257**	**0.0915**	**0.65**	[0.352, 0.082, 0.035]	[0.219, 0.022, 0.003]

Table 7: Llama3-70b Embedding Results

6 Analysis

Across our evaluation results discussed above, we observed that English color embeddings from all models evaluated to both higher fitness and correlation scores compared to the Chinese color embeddings, and significantly surpassed Japanese color embeddings. From our evaluation of Fugaku-LLM, we established that if it is indeed true that the model is trained on majority Japanese text, then this disparity cannot simply be attributed to the dominance of English in training data. We further hypothesize that a substantial number of Japanese color terms are English transliterations may be the reason, and recognize that Fugaku-LLM is undertrained compared to other models (billions of tokens instead of trillions).

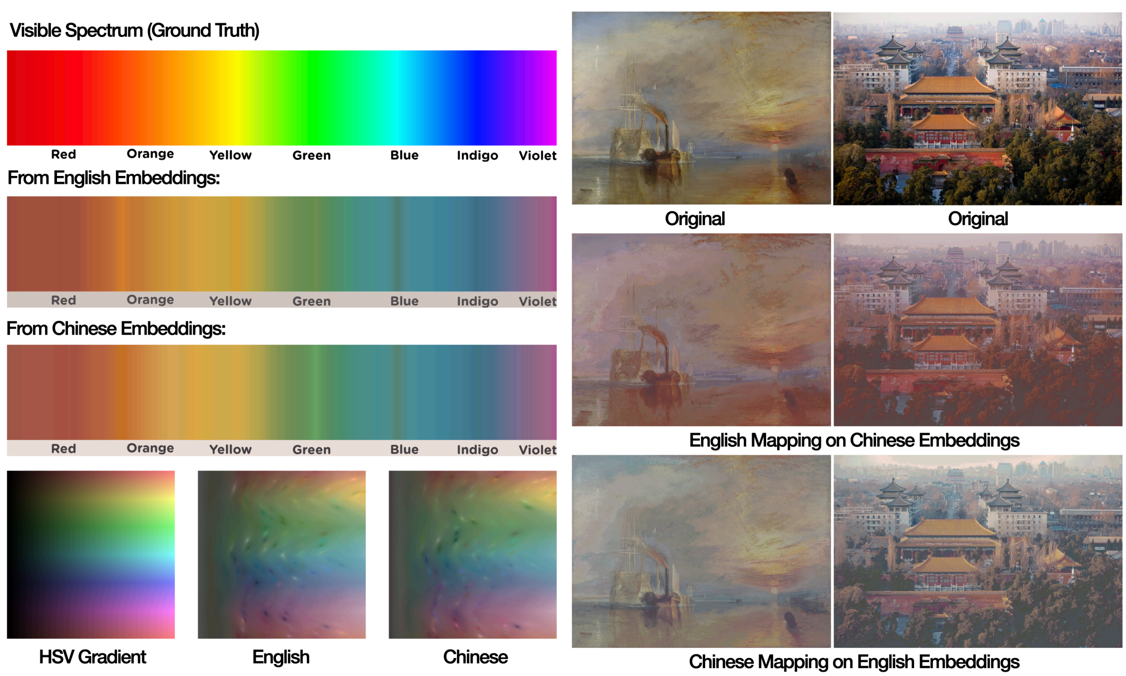

Figure 3: Result of linear mappings from embedding space to color space and when the mapping of one language is applied to another aligned embeddings; image credits: visible spectrum (Commons, 2024), HSV gradient (Commons, 2023b), *The Fighting Temeraire*, JMW Turner, (Commons, 2022), Beijing (Steinmetz)

Figure 3 shows select color mapping visualizations obtained (see Apppendix A.4). We observe that warmer colors are more accurately recovered from their representations in either English or Chinese. A greater range of green terms is recovered as either more yellow or blue hues, while cyan itself is mapped to a dark green. This finding is consistent with the observations made by Abdou et al. (2021) and supports the anthropological perspective on color naming articulated by Gibson et al. (2017): "Across languages, from the hunter-gatherer Tsimane' people of the Amazon to students in Boston, warm colors are communicated more efficiently than cool colors." Additionally, V-shaped banding artifacts are noticeable along lines of equal hue when the mapping is applied to a Hue-Saturation-Value (HSV) gradient, suggesting that the mapping primarily associates representations with specific hues, defined as the attributes of human visual perception that make an area seem similar to (a spectrum of) primary colors in a closed ring (Fairchild, 2013).

When a mapping originally fitted on aligned Chinese color embeddings is applied to aligned English embeddings, the resulting image features a cooler, more subdued palette, with more blue and gray tones, creating a subjectively more calm and serene atmosphere. Concomitantly, a mapping fitted on aligned English embeddings retrieves a more intense, warm palette from aligned Chinese embeddings, dominated by strong pinks, reds, and purples, creating a subjectively more dramatic and vibrant appearance.

Satisfactorily and precisely interpreting these visualized mapping results in linguistic and anthropological context is likely beyond the scope of this study. As hypothesized by Pavlick (2023), from a Conceptual Role Semantics (CRS) perspective, as LLM representations of color terms should be influenced by aspects conceptual role other than appearance despite our controlled context. As an example, the English mapping applied on Chinese color embeddings revealed that Chinese embeddings generally captured warmer color tones compared to English embeddings. A naive interpretation of this result may be that red is an auspicious color in Chinese culture, associated with vitality, celebration, and good fortune, such that red shows gain and green shows loss in Chinese stock exchanges and in many Asian countries, contrasting English usage where it is associated with more negative connotations. It could be said that a color transformation based on English interpretations of Chinese color terms produced a palette with a noticeable red shift, possibly reflective of a different emphasis or typical range of color usage in Chinese visual culture compared to English.

We additionally observe that the canonical correlation score for the "Red" component was consistently the highest, followed by "Green" and then "Blue". While this result could be explained by the evolutionary acquisition sequence of Basic Color Terms by cultures as conducted by the World Color Survey, where red precedes the the separation of green from blue (stage 1: "black", "white", stage 2: "red", stage 3: "blue-green" , stage 4: "yellow", and lastly distinction between "blue"and "green"in stage 5 (King, 2005)), we note that the randomly label-permuted embeddings also display this phenomenon, suggesting it stems from experimental setup. We believe this again highlights the importance of baseline comparisons in interpreting results.

7 Conclusion

Our investigation of Large Language Models' grounded representations extends a series of recently published studies by addressing gaps in multilingual analysis and baseline comparisons. The LLMs we examined consistently showed a better understanding of ground truths like color in English compared to in Chinese and Japanese, regardless of the model type and majority language in training data. However, improved alignment of color representations is achievable via targeted fine-tuning and model size increase. We additionally present a visualization for the qualitative analysis of varying cultural perceptions of the same underlying color space.

Our baselines, especially the GloVe baseline, suggest that traditional methods can also represent knowledge along the lines of symbolic structures defined by humans, and studies such as (Merullo et al., 2024) have begun to investigate the underlying mechanisms in LLMs. The reverse direction of integrating different modalities of data as embeddings is also exciting, with recent speculation that GPT-4o embeds images "directly into the transformer's semantic vector space" using a CNN architecture (Looney, 2024).

Our study's primary limitation is the constrained availability of cross-lingual color naming data, affecting the quality of our analysis as well as the accuracy of our visualizations that employ interpolation with inverse square falloff for unsampled input. The task of predicting color definitions from names is inherently challenging and ambiguous, as evidenced by the human baselines, and was chosen to diverge from (Abdou et al., 2021) by exploring a broader range of color terms instead of limiting the study to 51 monolexemic terms. For future work, we propose to augment our dataset and pool color naming data from geographically adjacent cultures, prompted by the observed improvements in Chinese color perception following finetuning in Japanese. Additionally, we aim to investigate the potential impacts of multimodal language models, focusing on how vision and language pretraining influence representations.

8 Ethics Statement

Given our project's focus on NLP and cross-cultural differences in color representation, we recognize the risk of cultural oversimplification. We use language models to infer about cultural color perceptions could inadvertently reinforce stereotypes or over-generalizations. Another risk is that our conclusions about the differences in cultural understanding of color may potentially be misused to advertise and exacerbate cultural division and intolerance. We are aware that the world color survey (Kay and Cook, 2014), as well as the Basic Color Terms theory proposed by Brent Berlin and Paul Kay, has been misused to advance a developmental hierarchy between cultures with biased "sociobiological and evolutionary-psychology implications" (Vox, 2017) (Saunders, 2000).

To address these concerns, we hereby emphasize the limitations and goals of our project. Our objective is to explore diverse cultural aspects to enhance cultural appreciation, not to provide a comprehensive overview of any culture. We have taken great care to maintain the cultural authenticity and accuracy of our dataset to avoid incorrect interpretations; both authors are bilingual in English and Chinese, with Author 1 additionally speaking Japanese. Our reproducible results from color mapping experiments should be viewed as only one of many possible ways to visualize the data and one of many potential insights, and our subjective analyses are limited to mood and aesthetics. We strictly intend these results for academic or cultural appreciation purposes and firmly oppose any misuse that supports discriminatory viewpoints.

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*, pages 109–132, Online. Association for Computational Linguistics.

AI@Meta. 2024. Llama 3 model card.

Guillaume Alain and Yoshua Bengio. 2016. Understanding intermediate layers using linear classifier probes. *ArXiv*, abs/1610.01644.

Morris Alper, Michael Fiman, and Hadar Averbuch-Elor. 2023. Is bert blind? exploring the effect of vision-and-language pretraining on visual language understanding.

Cyril Bailey. 1921. *On the Nature of Things. Translated by Cyril Bailey.* Oxford Clarendon Press.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR.

Wikimedia Commons. 2022. File:the fighting temeraire, jmw turner, national gallery.jpg — wikimedia commons, the free media repository.

Wikimedia Commons. 2023a. File:albert bierstadt - among the sierra nevada, california - google art project.jpg — wikimedia commons, the free media repository.

Wikimedia Commons. 2023b. File:hsv gradient example.png — wikimedia commons, the free media repository.

Wikimedia Commons. 2023c. File:lch gradient example.png — wikimedia commons, the free media repository.

Wikimedia Commons. 2023d. File:palace of westminster from the dome on methodist central hall (cropped).jpg — wikimedia commons, the free media repository.

Wikimedia Commons. 2024. File:colours of the visible light spectrum.png — wikimedia commons, the free media repository.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Harzen Europe. 2024. Ral color chart. Website. This site reviews standard colors according to the Classic RAL System, operated by Harzen Europe, an international reseller of genuine RAL products. It provides information defining standard colors for varnish, powder coating, and plastics, used widely in architecture, construction, industry, and road safety.

Mark D. Fairchild. 2013. *Color Appearance Terminology*, chapter 4. John Wiley & Sons, Ltd.

Maxwell Forbes, Ari Holtzman, and Yejin Choi. 2019. Do neural language representations learn physical commonsense?

Fujitsu. 2024. Release of "fugaku-llm" – a large language model trained on the supercomputer "fugaku". Fujitsu Global.

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):10785–10790.

Wes Gurnee and Max Tegmark. 2024. Language models represent space and time. In *The Twelfth International Conference on Learning Representations*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Toshiro Iijima, Wolfgang Wenning, and Heinrich Zollinger. 1982. Cultural factors of color naming in japanese: Naming tests with japanese children in japan and europe. *Anthropological Linguistics*, 24(2):245–262.

Aubrey G. Jaffer. Color dictionaries. `https://people.csail.mit.edu/jaffer/Color/Dictionaries`.

Genji Kawakita, Ariel Zeleznikow-Johnston, Naotsugu Tsuchiya, and Masafumi Oizumi. 2023. Comparing color similarity structures between humans and llms via unsupervised alignment.

Paul Kay and Richard S. Cook. 2014. *World Color Survey*. Springer Berlin Heidelberg, Berlin, Heidelberg.

Yova Kementchedjhieva, Sebastian Ruder, Ryan Cotterell, and Anders Søgaard. 2018. Generalizing Procrustes analysis for better bilingual dictionary induction. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 211–220, Brussels, Belgium. Association for Computational Linguistics.

Timothy D. King. 2005. Human color perception, cognition, and culture: why red is always red. In *IS&T/SPIE Electronic Imaging*.

koka. Irocore. `https://irocore.com`.

Nikolaus Kriegeskorte, Marieke Mur, and Peter Bandettini. 2008. Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2.

Martin Krzywinski. 2017. A large (9,284) list of named colors. `https://mk.bcgsc.ca/colornames/`.

Oran Looney. 2024. A picture is worth 170 tokens: How does gpt-4o encode images? https://www.oranlooney.com/post/gpt-cnn/.

Max M. Louwerse and Rolf A. Zwaan. 2009. Language encodes geographical information. *Cognitive Science*, 33(1):51–73.

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

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. 2018. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861.

Jack Merullo, Louis Castricato, Carsten Eickhoff, and Ellie Pavlick. 2023. Linearly mapping from image to text space.

Jack Merullo, Carsten Eickhoff, and Ellie Pavlick. 2024. Language models implement simple word2vec-style vector arithmetic.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013a. Exploiting similarities among languages for machine translation.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library.

Roma Patel and Ellie Pavlick. 2022. Mapping language models to grounded conceptual spaces. In *International Conference on Learning Representations*.

Ellie Pavlick. 2023. Symbols and grounding in large language models. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 381(2251):20220041.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations.

Alec Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pre-training.

Dave Raggett, Arnaud Le Hors, and Ian Jacobs. 1999. HTML 4.01 specification. W3C recommendation, W3C. Https://www.w3.org/TR/1999/REC-html401-19991224/.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA.

Anonymous Reviewers. 2024. Reviews for language models represent space and time. OpenReview. `https://openreview.net/forum?id=jE8xbmvFin`.

Sebastian Ruder. 2023. Do lms represent space and time? Newsletter. `https://newsletter.ruder.io/p/do-lms-represent-space-and-time`.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65:569–631.

Barbara Saunders. 2000. Revisiting basic color terms. *The Journal of the Royal Anthropological Institute*, 6(1):81–99.

Tal Schuster, Ori Ram, Regina Barzilay, and Amir Globerson. 2019. Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1599–1613, Minneapolis, Minnesota. Association for Computational Linguistics.

Navid Shokouhi. 2023. Canonical correlation analysis geometric interpretation. GitHub repository. https://github.com/idnavid/misc/blob/master/cca_geometricinterp.ipynb.

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, Melbourne, Australia. Association for Computational Linguistics.

George Steinmetz. Beijing-travel, nat geo image collection. https://www.nationalgeographic.com/travel/destination/beijing.

Mistral AI Team. Mistral-7b-instruct-v0.2. Hugging Face.

unsloth. 2024. Llama 3 8b bnb 4bit.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

Vox. 2017. The surprising pattern behind color names around the world. https://www.youtube.com/watch?v=gMqZR3pqMjg.

Yuxuan Wang, Wanxiang Che, Jiang Guo, Yijia Liu, and Ting Liu. 2019. Cross-lingual BERT transformation for zero-shot dependency parsing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5721–5727, Hong Kong, China. Association for Computational Linguistics.

Wikipedia contributors. 2023. Ral 颜色列表 — 维基百科, 自由的百科全书. https://zh.wikipedia.org/wiki/RAL .

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Jian-Qiao Zhu, Haijiang Yan, and Thomas L. Griffiths. 2024. Recovering mental representations from large language models with markov chain monte carlo.

Xiaowei Zhuang, Zhengshi Yang, and Dietmar Cordes. 2020. A technical review of canonical correlation analysis for neuroscience applications. *Hum Brain Mapp*, 41(13):3807–3833.

A Appendix (optional)

A.1 Human Color Prediction Baseline

The colors predicted by the two authors from bilingual color names as reported in the baselines in Table 3.

English	Chinese	Definition	Author 1	Author 2
Marble Grey	大理石灰			
Titanite Yellow	钛黄			
Jade Mussel Green	翡翠贻贝绿			
Deep Sea Blue	深海蓝			
Royal Purple	皇家紫			
Amethyst Light Violet	紫水晶浅紫			
Coral Pink	珊瑚粉			
Ember Red	灰烬红			
Mango Orange	芒果橙			
Rust Brown	锈棕			

A.2 Linear Mapping Performance

Example performance of regressor trained on Llama 3 8B:

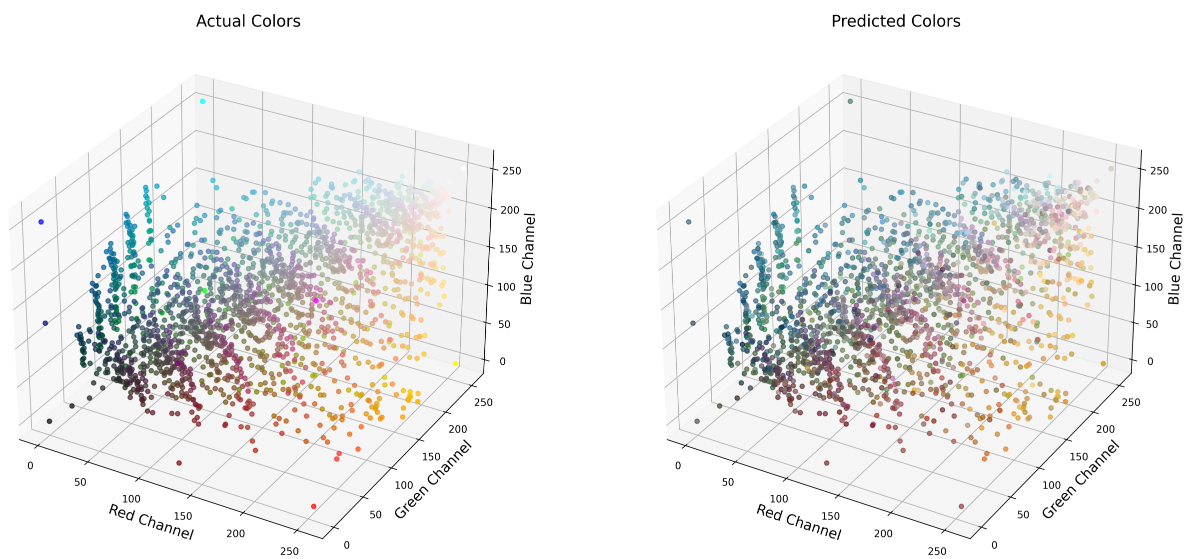

Figure 4: Examples from test set in RGB color space; points on the left are colored by actual definition; points on the right colored by prediction

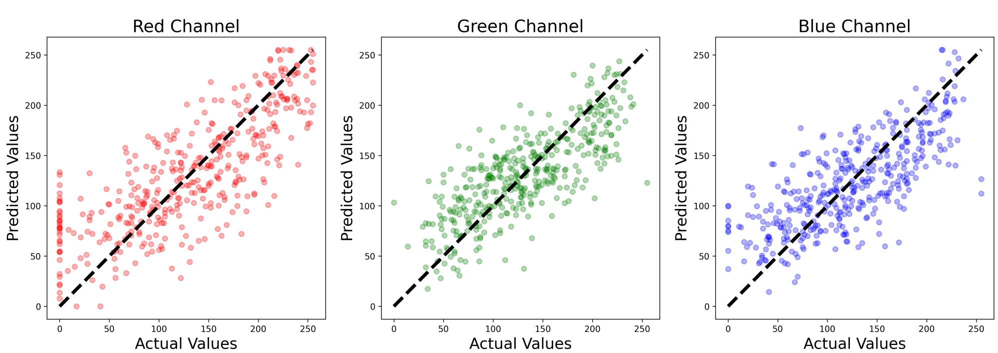

Figure 5: Regression performance on test set per channel

A.3 Finetuning Training Loss Graph

Figure 6: Loss over steps (3 epochs in total) of finetuning Llama3 8b 4bit

A.4 Additional Color Mapping Visualization

A.5 Chroma, Hue, Lightness (LCH) Color Gradient, (Commons, 2023c)

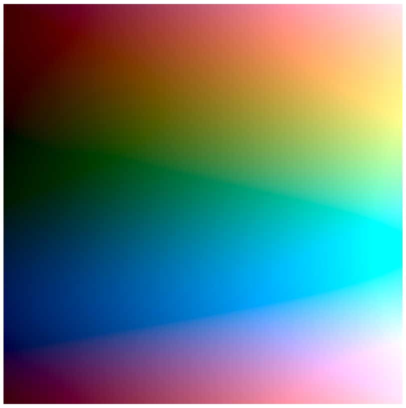

Figure 7: Chroma, Hue, Lightness (LCH) Color Gradient, (Commons, 2023c)

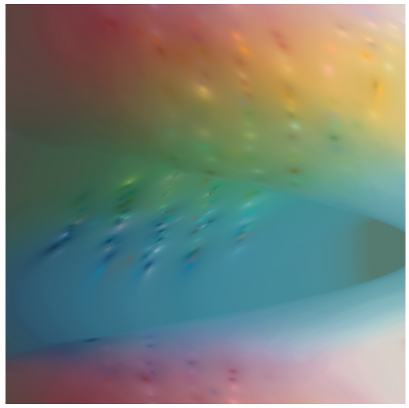

Figure 8: LCH Color Gradient, English Mapping of English Color Representations

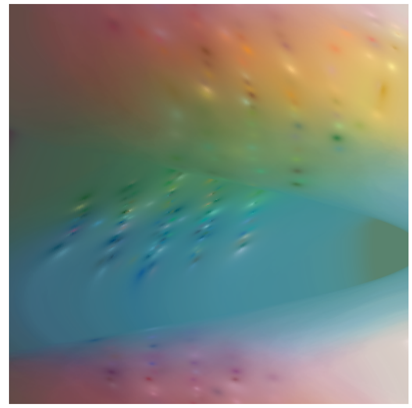

Figure 9: LCH Color Gradient, Chinese Mapping of Chinese Color Representations

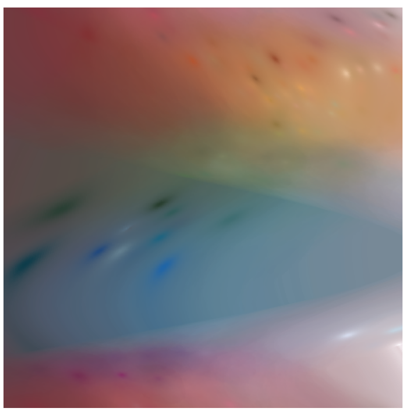

Figure 10: LCH Color Gradient, English Mapping of Chinese Color Representations

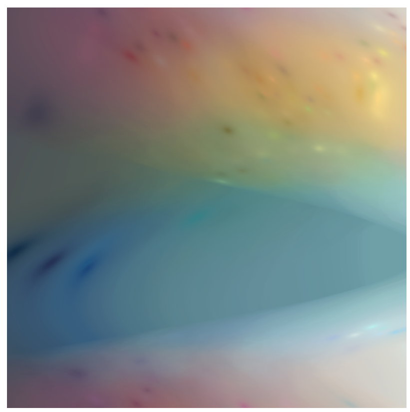

Figure 11: LCH Color Gradient, Chinese Mapping of English Color Representations

A.6 Palace of Westminster from the dome on Methodist Central Hall, (Commons, 2023d)

Figure 12: Palace of Westminster from the dome on Methodist Central Hall, (Commons, 2023d)

Figure 13: Photo of London, English Mapping of English Representations

Figure 14: Photo of London, Chinese Mapping of Chinese Representations

Figure 15: Photo of London, English Mapping of Chinese Representations

Figure 16: Photo of London, Chinese Mapping of English Representations

A.7 JMW Turner, *The Fighting Temeraire*, (Commons, 2022)

Figure 17: *The Fighting Temeraire*, JMW Turner

Figure 18: *The Fighting Temeraire*, JMW Turner, English Mapping of Japanese Color Representations

Figure 19: *The Fighting Temeraire*, JMW Turner, Japanese Mapping of English Color Representations

A.7.1 Albert Bierstadt, *Among the Sierra Nevada, California*

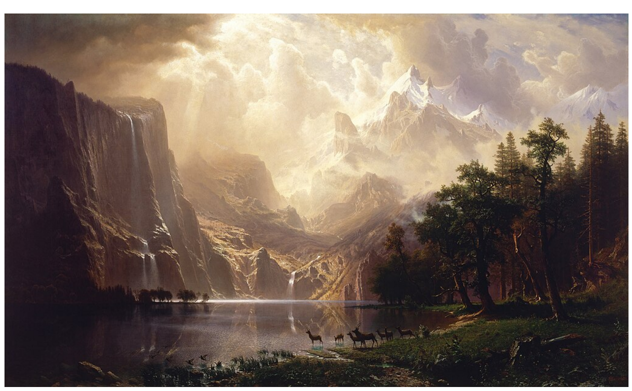

Figure 20: Albert Bierstadt, *Among the Sierra Nevada, California* (Commons, 2023a)

Figure 21: Albert Bierstadt, *Among the Sierra Nevada, California* (Commons, 2023a), English Mapping of English Color Representations

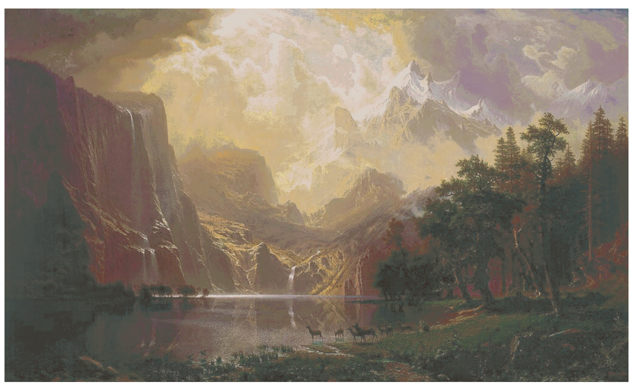

Figure 22: Albert Bierstadt, *Among the Sierra Nevada, California* (Commons, 2023a), Chinese Mapping of Chinese Color Representations

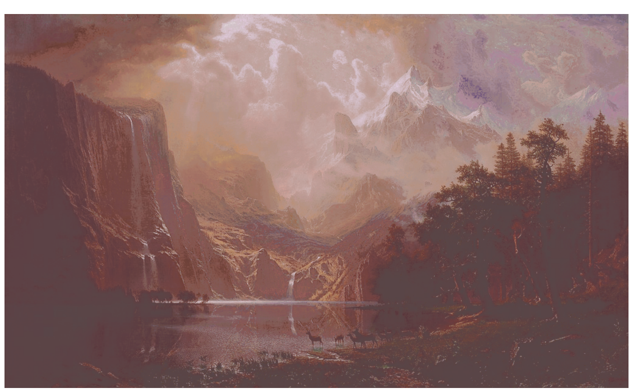

Figure 23: Albert Bierstadt, *Among the Sierra Nevada, California* (Commons, 2023a), English Mapping of Chinese Color Representations

Figure 24: Albert Bierstadt, *Among the Sierra Nevada, California* (Commons, 2023a), Chinese Mapping of English Color Representations