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

Pinlin [Calvin] Xu, Garbo Chung | {pinlinxu, garbo22}@stanford.edu

Overview

Are Large Language Models just stochastic parrots that recite the most probable answer, or do they learn representations of grounded structures in the real world to be able to answer your questions (as recently claimed)? We collected and built our own dataset of cross-lingual cultural color words, and perform a series controlled experiments to discover how the quality of representations varies with language, model, context, and fine-tuning. We also a color-mapping experiment that visualizes languages' color representation.

Data Collection & Preprocessing

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:

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

We also collected high-quality descriptions of 448 traditional Japanese colors detailing their composition, cultural impression and usage, etc.

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...

translation: "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 experiments and cultural analysis. Preprocessing is done to sanitize, deduplicate, remove nondescriptive names (such as Pantone "PMS 1485") and fix errors. For most experiments, a high-quality English-Chinese bilingual subset with 2054 examples is used.

Baselines

- **retrieval benchmark** for W3C webcolors: is the model a good stochastic parrot? **accurate on English; substantially worse on Chinese and Japanese**
- **randomly permuted** embeddings: correlate poorly with their original colors
- **human baselines** on color prediction from named entry: inherent difficulty
- **glove-wiki-gigaword-300**: baseline level of correlation $R^2 = 0.5952$

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, 0.05]
Human (Author 1)	0.5630	0.1356	N/A	N/A
Human (Author 2)	-0.0299	0.2219	N/A	N/A

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

Methods & Experiments

Large Language Models' Grounded Representations

We introduce three evaluation metrics of alignment between semantic embeddings of color terms and physics-based color coordinates in a specific color space:

- Regression analysis via probing[1]. We train a linear mapping with Elastic Net regularization and report out of sample fitness R^2

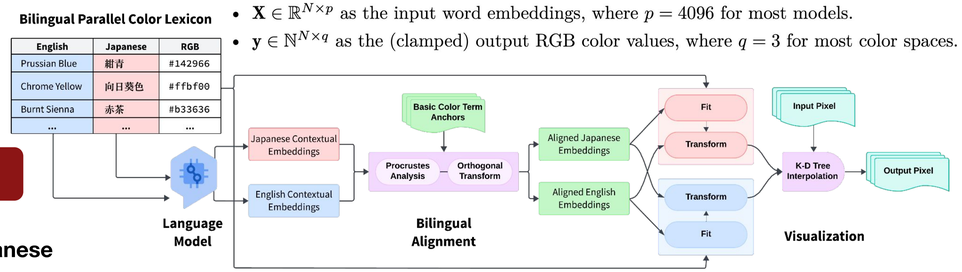
$$\mathcal{L}(\mathbf{W}, \mathbf{b}) = \frac{1}{N}\sum_{i=1}^{N} \|\mathbf{y}^{(i)} - \hat{\mathbf{y}}^{(i)}\|_2^2 + \lambda_1 \|\mathbf{W}\|_1 + \lambda_2 \|\mathbf{W}\|_2^2$$

- Canonical-Correlation Analysis (CCA). Dimensionality reduction is applied on overdetermined system using Principal Component Analysis (linear) or Uniform Manifold Approximation and Projection (nonlinear).

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

Color Mapping Visualization

- Motivated by previous research in cross-lingual embedding alignment, we introduce a pipeline to visualize differences in pairwise color representations. Given color term embeddings from a LLM in two languages, we align them to a joint space using orthogonal Procrustes analysis to optimally align the embeddings of Basic Color Terms, as identified by the World Color Survey[1]. Black, white, and red are used in final implementation.



- $\mathbf{X} \in \mathbb{R}^{N \times p}$ 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.

- 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, altering the perception and mood of an image when the color mapping is applied, though their interpretation is difficult and subtle.

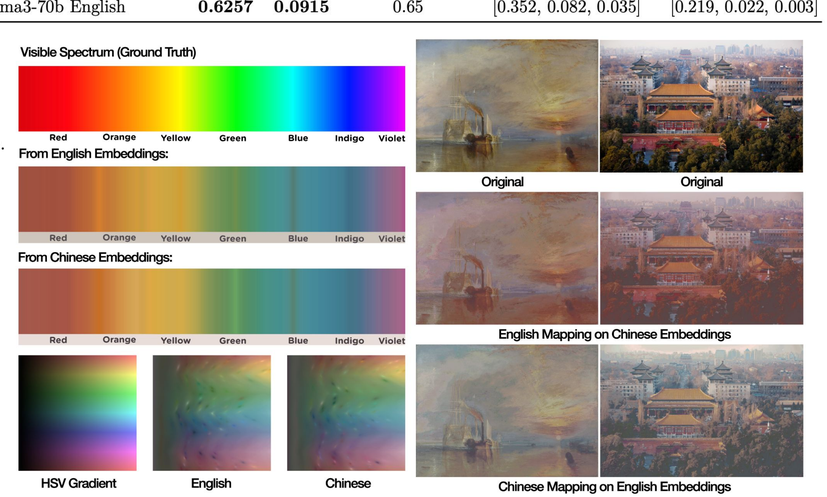
Discussion

- We found that across models, LLMs consistently demonstrates stronger representations for English color terms compared to Chinese and Japanese
- LLMs can achieve improved alignment for Chinese and Japanese embeddings through finetuning
- Finetuning on Japanese color dataset leads to a simultaneous increase in Chinese embedding alignment
- Alignment increases with model parameter size, outperforming GloVe
- A color transformation based on English mapping of aligned Chinese color embeddings produced a palette with a noticeable red shift, possibly reflective of differences in Chinese visual culture compared to English

Results

- If **LLMs exhibit a more grounded understanding of color in specific languages** (among English, Chinese, Japanese) → **Yes, English.**
- If **a model trained predominantly in a non-English language would possess superior representations in that language**: we analyzed the Fugaku-LLM-14B (60% training data in Japanese) → **No.**
- If **finetuning improves LLMs' color representation**: we finetuned Llama-3 on traditional Japanese colors. → **Yes, both Japanese & Chinese.**
- If **LLMs outperform GloVe baseline → No for small models (7B to 14B)**
- To assess if **representation improves with model size**: we analyzed English embeddings from Llama-3-70B. → **Yes**, surpassed glove-wiki-gigaword-300

Embeddings Configuration	R^2	d_{norm}	Weight Density	$\vec{\rho}_{PCA}$	$\vec{\rho}_{UMAP}$
Llama3 English	0.5544	0.0988	0.92	[0.488, 0.153, **0.127**]	[0.346, 0.217, 0.012]
Llama3 Chinese	0.4875	0.1054	0.92	[0.217, 0.181, 0.014]	[0.269, 0.130, 0.025]
Llama3 Japanese	0.2725	0.1325	**0.62**	[0.213, 0.143, 0.065]	[0.243, 0.089, 0.038]
Fugaku-LLM English	0.5994	0.0930	0.68	[0.329, 0.149, 0.055]	[0.317, 0.173, 0.020]
Fugaku-LLM Japanese	0.2701	0.1237	0.86	[0.158, 0.041, 0.002]	[0.229, 0.124, **0.060**]
Ft-Llama3 English	0.5544	0.0980	0.92	[**0.585**, 0.140, 0.101]	[**0.437**, 0.235, 0.037]
Ft-Llama3 Chinese	0.5049	0.1053	0.84	[0.299, 0.182, 0.001]	[0.196, 0.110, 0.001]
Ft-Llama3 Japanese	0.2813	0.1301	0.82	[0.238, **0.188**, 0.058]	[0.332, **0.260**, 0.041]
Llama3-70b English	**0.6257**	**0.0915**	0.65	[0.352, 0.082, 0.035]	[0.219, 0.022, 0.003]



- Warmer colors are more accurately recovered from their representations in either English or Chinese: "Across languages, from the hunter-gatherer Tsimane' people of the Amazon to students in Boston, warm colors are communicated more efficiently than cool colors" [3]
- V-shaped banding artifacts along lines of equal hue, defined as the attributes of human visual perception that make an area seem similar to (a spectrum of) primary colors in a closed ring

References

[1] Guillaume Alain and Yoshua Bengio. 2016. Understanding intermediate layers using linear classifier probes.ArXiv, abs/1610.01644.
[2] Paul Kay and Richard S. Cook. 2014. *World Color Survey*. Springer Berlin Heidelberg, Berlin, Heidelberg.
[3] 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.