Comparing Perceptual Judgements In Large Multimodal Models And Humans

Billy Dickson*[1], Sahaj Singh Maini *[1], Robert Nosofsky [2], Zoran Tiganj [1]

[1]Department of Computer Science, Indiana University Bloomington
[2]Department of Psychological and Brain Sciences, Indiana University Bloomington

INDIANA UNIVERSITY

Motivation

Evaluating and studying perceptual alignment between LMMs and Humans in decision making tasks.

Potential alternative for expensive human behavioral data collection.

Overview

Dataset

Existing dataset of 360 rock images with human ratings along 10 perceptual dimensions.

The dimensions were darkness/lightness, red/green, dull/shiny, chromaticity, smooth/rough, disorganized/organized, fine/coarse grain, porphyritic texture, conchoidal fractures and pegmatitic structure.

Models Tested

The study evaluated 4 LMMs – OpenAI GPT4, Anthropic Claude model family (Opus, Sonnet and Haiku).

Methodology

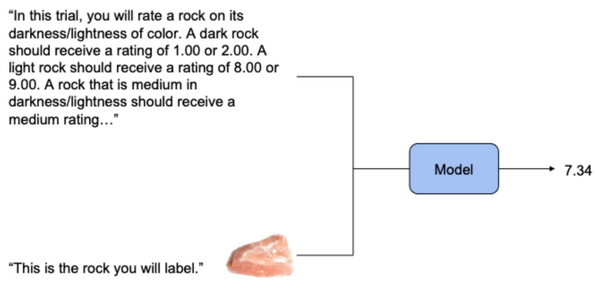
- Condition 1: Models were given verbal prompts without anchor images.
- Condition 2: Models were provided with verbal prompts and anchor images.
- Condition 3: Exploratory attempts to improve performance on challenging dimensions using detailed prompts.

Key Findings

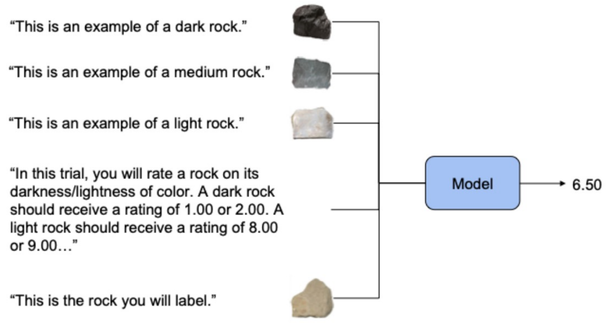
- GPT-4 and Claude-3 Sonnet performed best overall .
- Models performed better on elementary visual dimensions (e.g., lightness, color) than on abstract or emergent dimensions (e.g., organization, pegmatitic structure).
- Adding anchor images in context generally improved performance

Prompt Structure

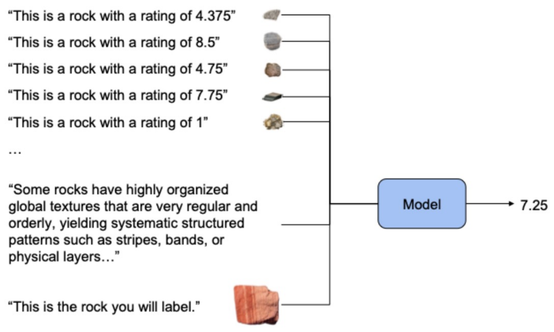
Condition 1



Condition 2



Condition 3



Interactive Plots

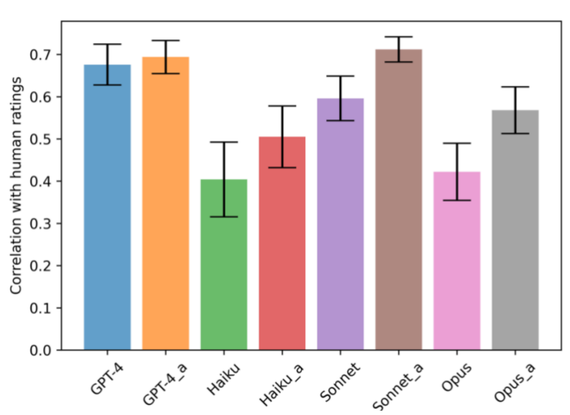
Paper

Results

Correlations Between Models And Humans In Each Dimension

dimension/model	GPT4	GPT4 with anchor	Haiku	Haiku with anchor	Sonnet	Sonnet with anchor	Opus	Opus with anchor
chromaticity	0.77	0.80	0.81	**0.82**	0.80	0.77	0.72	0.80
darkness/lightness	0.88	0.85	0.86	0.87	0.84	**0.89**	0.79	0.83
disorganized/organized	0.48	0.42	0.25	0.12	0.45	**0.57**	0.28	0.33
dull/shiny	**0.81**	0.72	0.40	0.48	0.75	0.61	0.33	0.47
fine/coarse grain	0.76	**0.80**	-0.02	0.53	0.61	0.76	0.25	0.46
red/green	0.82	0.78	0.53	0.71	0.67	**0.83**	0.59	0.82
smooth/rough	0.48	0.63	0.04	0.24	0.32	**0.68**	0.19	0.42
conchoidal fracture	0.67	**0.71**	0.30	0.53	0.55	0.70	0.29	0.43
pegmatitic structure	0.43	0.56	0.28	0.32	0.38	**0.66**	0.20	0.57
porphyritic texture	0.66	**0.67**	0.59	0.43	0.59	0.65	0.58	0.55

Mean And Standard Deviation Across Conditions



References

Meagher, B. J., & Nosofsky, R. M. (2023). Testing formal cognitive models of classification and old-new recognition in a real-world high-dimensional category domain. Cognitive Psychology, 145, 101596.

Nosofsky, R. M., Sanders, C. A., Meagher, B. J., & Douglas, B. J. (2018). Toward the development of a feature-space representation for a complex natural category domain. Behavior Research Methods, 50(2), 530–556.

Nosofsky, R. M., Sanders, C. A., Meagher, B. J., & Douglas, B. J. (2020). Search for the Missing Dimensions: Building a Feature-Space Representation for a Natural-Science Category Domain. Computational Brain & Behavior, 3(1), 13–33.

Nosofsky, R. M., Sanders, C. A., Meagher, B. J., & Douglas, B. J. (2018). Toward the development of a feature-space representation for a complex natural category domain. Behavior Research Methods, 50(2), 530–556.

Nosofsky, R. M., Sanders, C. A., Meagher, B. J., & Douglas, B. J. (2020). Search for the Missing Dimensions: Building a Feature-Space Representation for a Natural-Science Category Domain. Computational Brain & Behavior, 3(1), 13–33.