Grounded Concreteness : Human-Like Concreteness Sensitivity in Vision-Language Models

Aryan Roy, Zekun Wang, Christopher J. MacLellan
{aroy389, zekun, cmaclell}@gatech.edu
School of Interactive Computing, Georgia Institute of Technology

Motivation

- LLMs learn primarily from text, no interaction with the physical world, psychological evidence points to **dual-coding**
- Hard to gauge alignment between models and humans through result driven benchmarks
- If meaning is dependent on grounding, then how do we measure it?

Concreteness

- We operationalize concreteness using a psycholinguistic measure of how much a concept refers to a perceptible, tangible entity. Human norms from **Brysbaert et al., 2014**
- **Sentence level concreteness = mean token concreteness**

$$C(x) = \frac{1}{n}\sum_{i=1}^{n} c(w_i).$$

Does Vision Pre-Training Impact How Models Handle Concreteness?

Dataset	Domain
Arc-Easy	Grade-School Science
Arc-Challenge	Grade-School Science (Hard)
BoolQ	Reading Comprehension
WinoGrande	Commonsense (Coreference)
CommonsenseQA	General Commonsense
PIQA	Physical Interaction
SIQA	Social Commonsense

VLMs offer an opportunity to do a controlled ablation and study the effect of concreteness of the prompt.

We picked two **non-instruct LLM-VLM pairs** with a **common language backbone** and ran inference on **text-only datasets**.

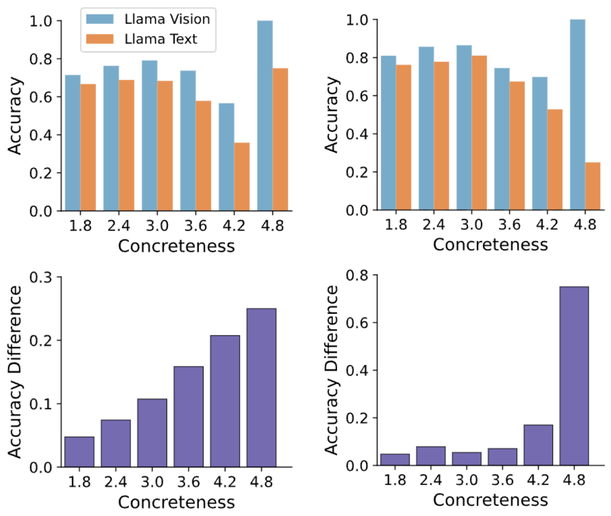
	Model Family	Size	# of Layers
Small Pair	Llama 3.1 (Text)	8B	32
	Llama 3.2 (Vision)	11B	32
Large Pair	Llama 3.1 (Text)	70B	64
	Llama 3.2 (Vision)	90B	64

Llama 3 Model family

Behavioral Differences

The **difference in accuracy grows with concreteness** ; monotonically for small models

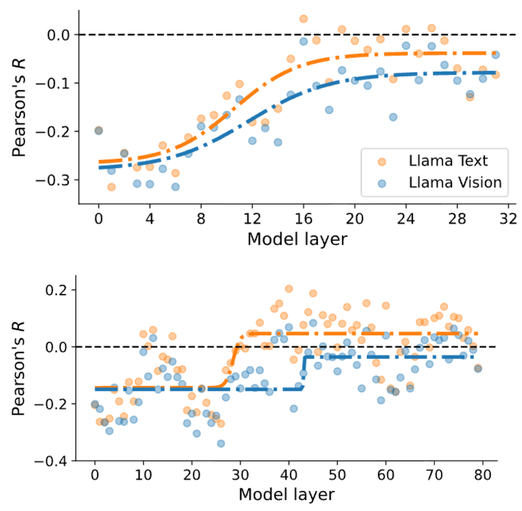
Small Pair **Large Pair**



VLMs reason better with prompts containing more concrete tokens

Attention Entropy

Abstract Words depend on neighboring tokens to derive meaning. We measured the layer wise **Pearson's r** between token **concreteness** and **head-averaged attention entropy**.



Small Pair

Large Pair

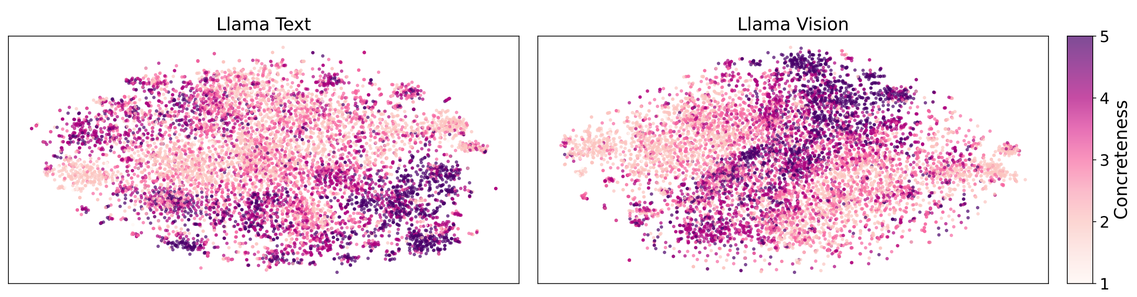
Concrete tokens exhibit lower attention entropy.

Representation Geometry

Any meaningful difference in behavior based on concreteness should be reflected in the **representation geometry**.

Visual grounding encourages representations of perceptually **grounded words to occupy a more coherent subregion** of the space!

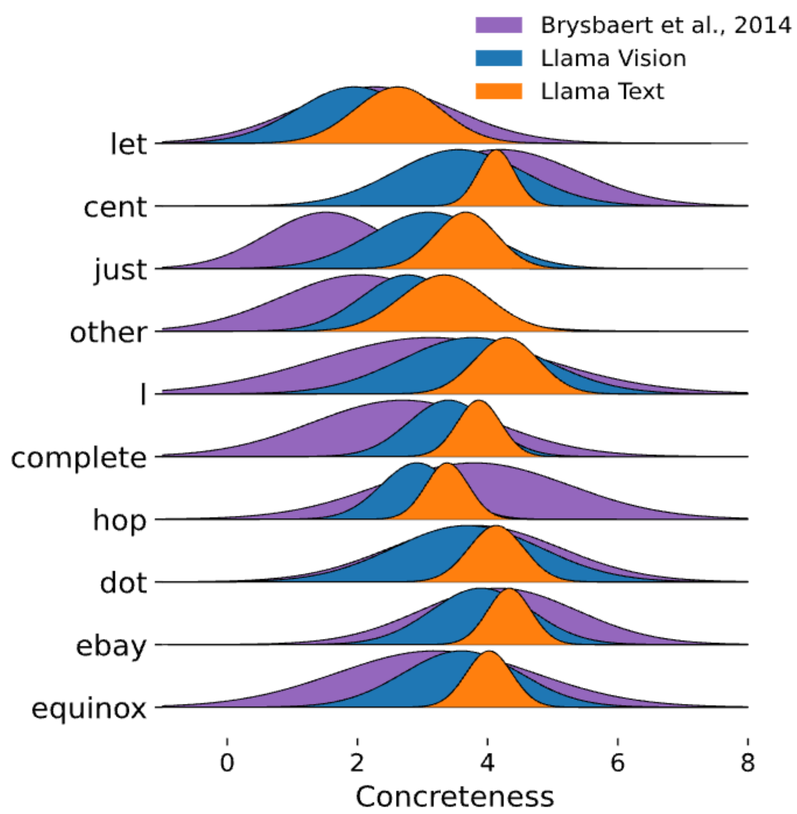
Small Pair



Large Pair

Model concreteness behavior carries through scale!

LLM Judgement Vs Human Norms



We also measured alignment between LLM Judgement and Human norms using KL Divergence.

VLM exhibits lower divergence than the LLM mean DKL: 9.4 vs. 10.1.