2736

Evaluating Model Alignment with Human Perception: A Study on *Shitsukan* in LLMs and LVLMs

shitsukan-eval

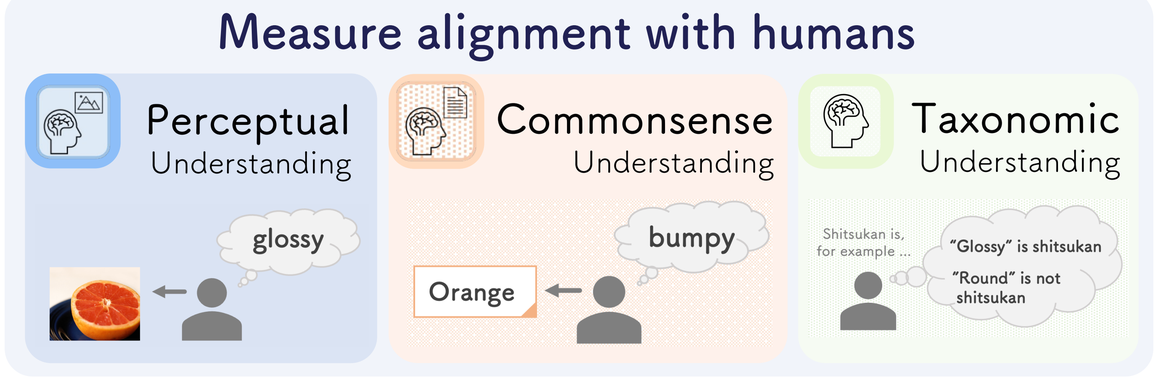
Daiki Shiono[1] Ana Brassard[2,1] Yukiko Ishizuki[1,2] Jun Suzuki[1,2,3]
[1]Tohoku University [2]RIKEN [3]National Institute of Informatics

Abstract

- We focus on the concept of *"Shitsukan"*, which encompasses the entire sensory experience when perceiving objects
- This study aims to verify how much alignment there is between human judgment and LLMs/LVLMs regarding *Shitsukan*
- We constructed **datasets** of *Shitsukan* terms and **benchmarks** that evaluated the *Shitsukan* recognition ability of LLMs/LVLMs

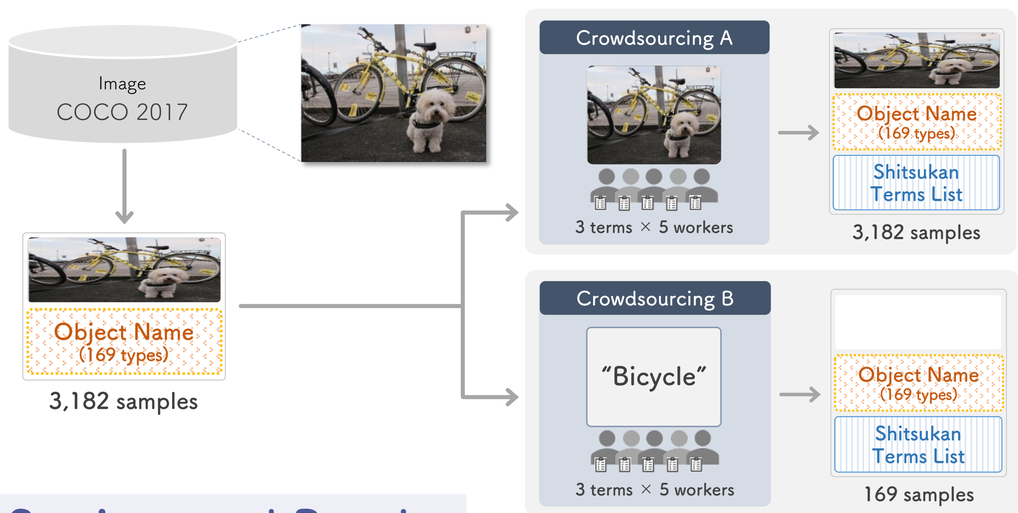
Background: The Concept of *Shitsukan*

- *Shitsukan* : Expressions that include the physical properties, states, and impressions of objects
- **Goal:** Investigating the extent to which LLMs/LVLMs can capture and reflect *Shitsukan* characterized by human experience



Dataset Construction

- Investigating how humans recall *Shitsukan* terms for certain objects
- Created common objects in context dataset (COCO) with *Shitsukan*



Experiments

Perceptual Understanding

Q: Is this *shitsukan* natural for a certain object image?

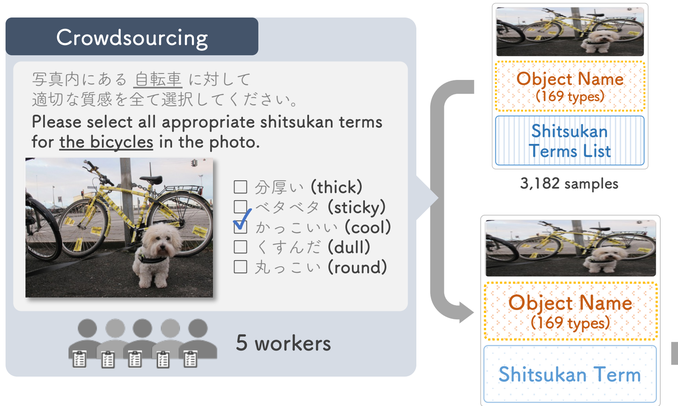
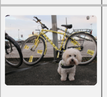


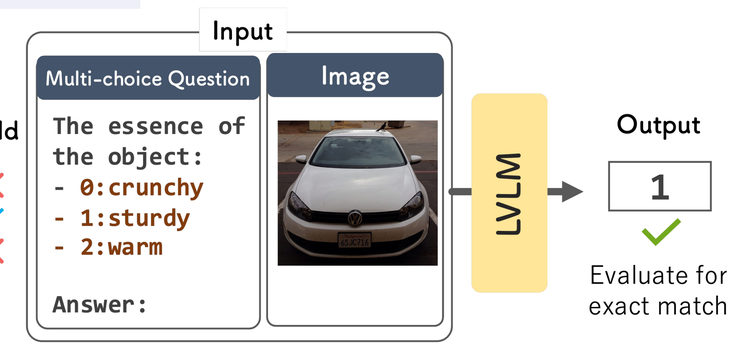
Image	Object	POS 5 votes	POS 4 votes	3 votes	2 votes	NEG 1 votes	NEG 0 votes
	自転車 (bicycle)	止まっている (stationary)		かっこいい (cool)		微光沢 (low sheen)	乾燥した (dry) ゴツゴツ (rugged)

assigned through a majority vote

- Check the appropriateness level of the collected *Shitsukan* terms
- Determining if a pair of { Image , Shitsukan Term } is appropriate
- To use in the Taxonomic Understanding task, aggregating into POS(*Shitsukan* terms) / NEG(Non-*Shitsukan* terms)

Settings and Results

- Created multi-choice questions to be answered based on an image object

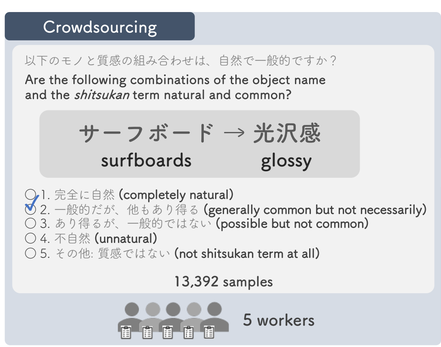


	Model	Number of Choices 2	3	4	5
Ja	GPT-4V	88.4	81.2	77.6	74.6
	Qwen2-VL$_{7B}$	85.4	72.5	64.8	60.9
En	GPT-4V	88.4	77.3	70.4	63.9
	Qwen2-VL$_{7B}$	80.0	69.6	64.5	57.3

The newest models, Qwen2-VL$_{7B}$, GPT-4V consistently performed better in Japanese than in English

Commonsense Knowledge

Q: Is this *shitsukan* natural for a certain object?



Object	1 completely natural	2 generally common but not necessarily	3 possible but not common	4 unnatural	5 not shitsukan term at all
サーフボード (surfboards)	固い (hard) 硬い (stiff)	なめらか (smooth) 軽そう (looks light) 光沢感 (shiny)	大きい (big) かさかさ (dry) ザラザラ (rough)	涼しい (cool) ゴツゴツ (rugged)	軽やかで (light) 整頓 (tidy)

assigned through a majority vote

- Evaluating the naturalness of a pair of { Object Name , Shitsukan Term } on a 5-point scale

Settings and Results

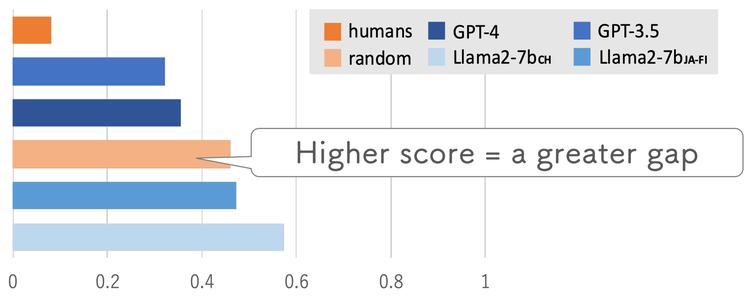
- Calculated the distribution of labels for each { Object Name , Label List }$_i$ pair :

$$f_i(x) = \frac{1}{\sqrt{2\pi\sigma^2}} exp(-\frac{(x-\mu)^2}{2\sigma^2})$$

(μ: mean, σ: standard deviation, $i \in \{1, 2, \ldots, N\}$)

- Calculated the average value of distance:

$$d_i = |f_i(l_{most\_freq}) - f_i(l_{pred})| \quad (0 < d_i < 1)$$



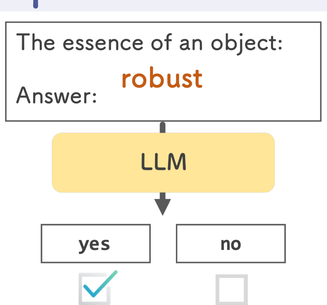
There remain significant differences between gold and the current LLMs

Taxonomic Understanding

Q: Which terms are appropriate as *Shitsukan* ?

Settings and Results

- Generating yes/no if the given object name is appropriate as shitsukan



	Model	Acc (%)		Model	Acc (%)
Ja	GPT-3.5	65.7	En	GPT-3.5	65.5
	GPT-4	74.3		GPT-4	74.8
	Llama2-7b$_{JA-FI}$	65.6		Llama2-7b$_{CH}$	78.3

* Llama2-7b$_{JA-FI}$: ELYZA-japanese-Llama-2-7b-fast-instruct
* Llama2-7b$_{CH}$: Llama-2-7b-hf-chat

Current LLMs can answer whether a term is a *shitsukan* term or not with less than 80% correct responses