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

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
{yukiko.ishizuki.p7, daiki.shiono.s1}@dc.tohoku.ac.jp
ana.brassard@riken.jp jun.suzuki@tohoku.ac.jp

Abstract

We evaluate the alignment of large language models (LLMs) and large vision-language models (LVLMs) with human perception, focusing on the Japanese concept of *shitsukan*, which reflects the sensory experience of perceiving objects. We created a dataset of *shitsukan* terms elicited from individuals in response to object images. With it, we designed benchmark tasks for three dimensions of understanding *shitsukan*: (1) accurate perception in object images, (2) commonsense knowledge of typical *shitsukan* terms for objects, and (3) distinction of valid *shitsukan* terms. Models demonstrated mixed accuracy across benchmark tasks, with limited overlap between model- and human-generated terms. However, manual evaluations revealed that the model-generated terms were still natural to humans. This work identifies gaps in culture-specific understanding and contributes to aligning models with human sensory perception. We publicly release the dataset to encourage further research in this area.

cl-tohoku/shitsukan-eval

1 Introduction

Ensuring that large language models (LLMs) and large vision-language models (LVLMs) share the same understanding of the world as humans is essential for their utility in real-world applications, where models must perceive, act, and communicate in ways aligned with human understanding. This work focuses on *perception*, a complex process involving the multi-sensory experience of objects (Fleming et al., 2015a), which humans often describe through language. These descriptions are often vague, relying on subjective expressions of “feelings” and “impressions.” We examine whether models can capture and reflect these nuanced human experiences.

As a case study, we examine the Japanese concept of *shitsukan*, which captures the sensory experience when perceiving objects (Spence, 2020).

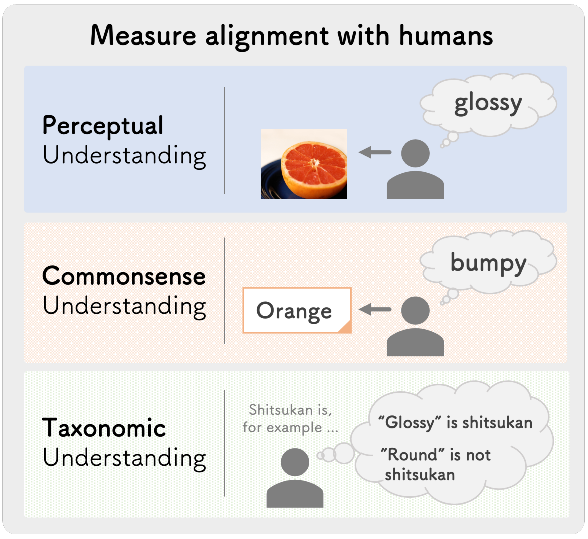


Figure 1: We evaluated LLM/LVLMs’ understanding of *shitsukan* using newly-built datasets. We measured three dimensions of understanding: perception, commonsense knowledge, and taxonomic knowledge.

Native Japanese speakers intuitively understand and express *shitsukan* when prompted, despite finding it difficult to define formally. Its boundaries are blurry and highly subjective. A sense of *shitsukan* is not explicitly taught but rather acquired through life experiences, context, and observing others (Komatsu and Goda, 2018). This makes it an ideal subject for assessing how well models can handle vague sensory experiences. Additionally, *shitsukan* is a culturally unique concept that is difficult to translate into English, broadening the appeal of this study as a step away from Western-centric datasets and tasks.

In this work, we present new datasets of *shitsukan* terms elicited from individuals in response to object images and object names: a crowdsourced dataset consisting of 26,223 instances of {image, object, *shitsukan* term} triples (§3), and additional data that we used to build benchmark tasks. Using the new data, we analyzed the current capabilities

Proceedings of the 31st International Conference on Computational Linguistics, pages 11428–11444
January 19–24, 2025. ©2025 Association for Computational Linguistics

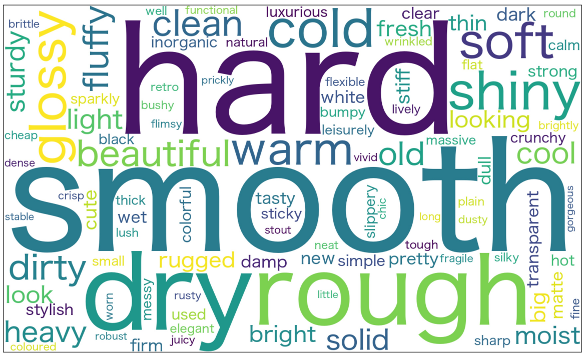


Figure 2: Word cloud of the most frequent English *shitsukan* terms in the dataset.

of LLMs and LVLMs in recognizing *shitsukan*. As illustrated in Figure 1, we considered three dimensions of understanding for models:

- ***Perceptual***: Accurately describing *shitsukan* in object images (§4);
- ***Commonsense***: Having accurate knowledge of appropriate terms typically associated with given objects (§5);
- ***Taxonomic***: Distinguishing which terms qualify as *shitsukan* (§6).

While some models demonstrated limited capabilities, and the overlap between model-generated and human-written terms was low, manual evaluations revealed that models produced *shitsukan* terms that appeared natural to human observers. We release the dataset, including a full English translation, to support further research in this area.

2 Background

2.1 *Shitsukan*

Shitsukan (JA: 質感) literally translates to "essential qualities." Colloquially, it describes the multi-sensory experience when perceiving an object, encompassing its texture, material, or general impression. The concept is best illustrated through examples, such as the word cloud of common terms in our dataset shown in Figure 2, with its Japanese counterpart provided in Figure 7 in the appendix.

Formally, Shinmura (2018) defined *shitsukan* as: "*The feeling you get from the difference in the nature of the material*" and "*The feeling that the material originally has.*" In research contexts, it is considered to include (1) the essential qualities of an object and (2) the psychological functions triggered by perceiving sensory stimuli (Komatsu and Goda, 2018). In this study, we adopt the following definition:

> *Shitsukan* terms are expressions that include the ***physical properties***, ***states***, and ***impressions*** of objects.

While it aligns with previous works (e.g. Komatsu and Goda, 2018), the definition is intentionally simple and concrete, aiming to balance intuitive understanding with clarity to aid annotation by non-experts. Notably, while closely related, this definition is broader than the English concept of *texture*, expanding it to include not only tactile properties but also visual, auditory, and other sensory attributes.

2.2 Related Studies

Shitsukan has been explored across various fields, including neuroscience, psychology, and linguistics. In neuroscience, studies have focused on identifying neural pathways involved in texture perception (Wada et al., 2014; Komatsu and Goda, 2018). Psychophysics research has examined its common dimensions, e.g., roughness/smoothness, hardness/softness, and coldness/warmness (Okamoto et al., 2013). *Shitsukan* is recognized as a fundamental and ubiquitous concept in human visual cognition research (Adelson, 2001; Maloney and Brainard, 2010; Fleming et al., 2015b), with extensive studies on object characteristics that evoke *shitsukan* (Nishida and Shinya, 1998; Motoyoshi et al., 2007; Marlow et al., 2012). In psychology and linguistics, phonetic correlations between words and textures have been observed across languages (Maurer et al., 2006; Sakamoto and Watanabe, 2018; Winter et al., 2022), highlighting the deep entanglement of language and sensory perception. *Shitsukan* has been studied in Japanese using small sets of typical terms (Nishinari et al., 2008; Hayakawa et al., 2013), but there is no comprehensive dataset capturing how humans recall *shitsukan* terms for specific objects.

In turn, in natural language processing (NLP), some research has focused on developing models that mimic human cognition, such as commonsense reasoning, resulting in numerous benchmarks (Talmor et al., 2019; bench authors, 2023, *i.a.*). Some benchmarks also address more complex and subjective concepts, like humor (Amin and Burghardt, 2020; Hessel et al., 2023). Closer to *shitsukan* and world grounding, studies have investigated how language models implicitly capture the concept of color (Abdou et al., 2021; Paik et al., 2021). Others have also explored the concept of texture, which falls under the broader framework of *shitsukan* (Cimpoi et al., 2014; Wu et al., 2020). Additionally, there has been a significant increase in research focusing on multi-modal models considering various aspects of perception (Liu et al., 2023; Schwenk et al., 2022; Ye et al., 2023). However, unlike in neuroscience and psychology, the concept of "*shitsukan*" has not received much attention, and no data exists that could be used to evaluate models' understanding.

3 Dataset Construction

We constructed a dataset of *shitsukan* terms recalled by individuals in response to object images, then used it to design test sets for LLM/LVLM benchmarking. This section outlines the main data collection process, while the benchmark constructions are detailed in their respective experiment setting descriptions (§4.1, §5.1, §6.1).

3.1 Source Data

We began by sampling representative images of objects in various states and everyday contexts from the Common Objects in Context (COCO) dataset (Lin et al., 2014), specifically the 2017 train set. To ensure quality and safety, we excluded images with the "other" tag due to insufficient object information, as well as those tagged as "bathroom" or "person" to avoid inappropriate or personal content. This filtering left 169 `object` and `stuff` tags, which we translated into Japanese. We further excluded images in which the target object's bounding box occupied less than 15% of the image area to ensure clear visibility of the object. Finally, we randomly sampled up to 20 images per object tag,[1] resulting in a total of 3,182 images to be annotated with *shitsukan* terms.

3.2 Collection Task

Annotators were presented with an image of an object, the object's name, and a free-text field to provide three *shitsukan* terms for the object (Figure 3). We also provided a definition of *shitsukan*, as outlined in Section 2.1, with formatting instructions and examples for clarity.

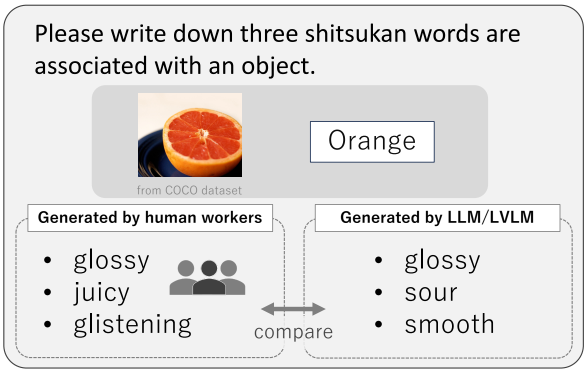


Figure 3: We created a *shitsukan* dataset by collecting terms recalled by individuals in response to images and object names.

3.3 Crowdsourcing Protocol

Crowdsourcing was conducted on the Yahoo! Crowdsourcing platform.[2] Each object-image pair was annotated by three different workers. Workers were strictly screened through several pilot rounds and manually selected based on their performance on trial tasks. During the main collection rounds, annotators' work was continuously reviewed, and those failing to adhere to the guidelines were disqualified. Their contributions were excluded, and the corresponding data was recollected in subsequent rounds. Further details on worker selection, compensation, example forms, and instruction screens are provided in Appendix A.1.

3.4 Data Cleaning

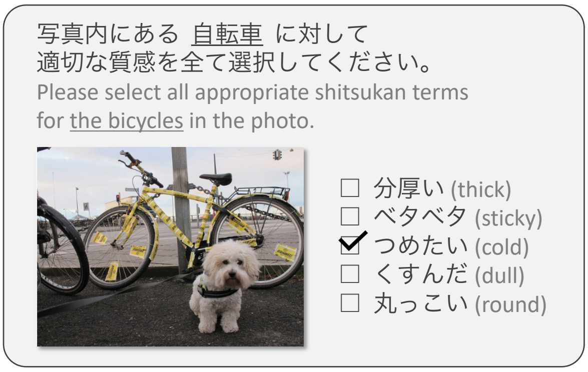
We also conducted a series of data-cleaning steps, removing responses that did not meet the formatting requirements or included invalid expressions such as phrases or sentences. We re-collected data until all images had a minimum of six valid terms. The dataset underwent additional cleaning during the translation to English, described in the next section. In total, we collected 22,409 Japanese *shitsukan* terms, of which 2,665 are unique, each paired with an image and object tag.

3.5 Translation to English

Although our primary focus is on Japanese data, we translated the dataset to English to enable broader usage and comparison with existing datasets. The translation process involved automatic translation followed by manual verification by professional translators.

[1] Some object tags had fewer than 20 candidate images.

[2] `https://crowdsourcing.yahoo.co.jp/`



写真内にある 自転車 に対して
適切な質感を全て選択してください。
Please select all appropriate shitsukan terms
for the bicycles in the photo.

☐ 分厚い (thick)
☐ ベタベタ (sticky)
☑ つめたい (cold)
☐ くすんだ (dull)
☐ 丸っこい (round)

Figure 4: Example of the *shitsukan* selection task (§4.1). The task is to select all terms that describe the object in the image. English translations are added for clarity.

Automatic translation was conducted using DeepL[3]. To encourage using the correct word sense, we added the context of its associated object using the template "This {object} is {*shitsukan*}."[4]

During manual verification, translators checked the accuracy of the Japanese-to-English translations within the context of the associated images. They flagged terms that clearly referred to a different object, noting the appropriate object instead (either an existing or new tag). Translators also ensured that translated terms aligned with the concept of *shitsukan*, removing any completely invalid terms. In cases where multiple Japanese terms mapped to the same English term, translators were encouraged to refine translations for specificity (e.g., distinguishing ふわふわ (fuwafuwa) and もふもふ (mofumofu), both commonly translated as "fluffy"). As a result, 25.4% of the *shitsukan* term translations were manually corrected. The final English dataset contains 1,240 unique terms.

4 Perception

Our first set of experiments addresses *shitsukan* perception, i.e., ***whether models can recognize shitsukan in images*** as humans do. Specifically, we reproduced the annotation process with LVLMs and compared the generated terms with the original human responses. Additionally, we prepared an alternative classification setting where models were tasked to *select* the most appropriate term.

4.1 Task Design

Generation. Given the same information as in the original crowdsourcing (target image, object tag, instructions, examples; §3.3), the task is to generate three *shitsukan* terms that best describe the specified object in the image.

Selection. Given a target image, object tag, and a list of *shitsukan* terms, the task is to select the most appropriate term for the specified object in the given image (Figure 4). We compared performance in multiple-choice settings, with one correct answer and one to four alternatives. Each selection task (binary, three-way, four-way, and five-way) included 335 samples in Japanese and English, respectively. Positive and negative samples were determined by the number of votes collected from human annotators, as described in the following section.

4.2 Selection Benchmark Construction

We conducted additional crowdsourcing, in which annotators were asked to select one or more terms that applied to the specified object in a given image. Each task presented five candidate terms: three sampled from the terms originally provided by workers for that image and object and two randomly sampled from terms associated with other objects. Five workers annotated each form. Further details on crowdsourcing can be found in Appendix A.2. After data collection, terms were tagged as positive samples if at least four annotators selected them and as negative samples if one or no annotator selected them. This process also served as a validation step, as invalid terms were unlikely to be selected, and as a discovery step, allowing for identifying new valid combinations when randomly paired terms were selected as acceptable. Finally, each classification test sample was constructed by pairing one positive sample term with one or more negative sample terms.

4.3 Experimental Settings

Models. We evaluated several open-source and proprietary models, including from the Qwen2-VL (Wang et al., 2024), Llama 3-V (Dubey et al., 2024), LLaVA-OneVision (Li et al., 2024), LLaVA-NeXT (Liu et al., 2024b), LLaVA-1.5 (Liu et al., 2024a), and GPT-4 (OpenAI, 2023) families. At the time of writing, these models were reported to have strong multimodal and multilingual performance in their latest technical reports. An example

[3] `https://www.deepl.com/translator`

[4] Originally: "この{object}は{*shitsukan*}"

	Model Name	Number of Candidates			
		2	3	4	5
	Chance rate	50.0	33.3	25.0	20.0
JA	Qwen2-VL$_{7B}$	85.4	72.5	64.8	60.9
	Llama-3.2-Vision$_{11B}$	57.0	62.1	49.9	45.1
	LLaVA-OneVision$_{7B}$	79.4	77.9	66.3	63.0
	LLaVA-NeXT$_{7B}$	55.5	38.5	33.4	23.9
	LLaVA-1.5$_{7B}$	52.5	32.2	26.3	21.8
	GPT-4o-20241120	93.7	87.2	85.4	80.9
	GPT-4V-20240409	88.4	81.2	77.6	74.6
	GPT-4V*	87.5	79.1	76.1	74.0
EN	Qwen2-VL$_{7B}$	80.0	69.6	64.5	57.3
	Llama-3.2-Vision$_{11B}$	83.6	70.1	63.6	54.6
	LLaVA-OneVision$_{7B}$	81.8	76.1	73.1	64.8
	LLaVA-NeXT$_{7B}$	60.6	49.3	37.9	34.0
	LLaVA-1.5$_{7B}$	61.2	49.6	39.7	34.6
	GPT-4o-20241120	87.5	78.5	76.4	74.3
	GPT-4V-20240409	88.4	77.3	70.4	63.9
	GPT-4V*	53.4	43.9	52.8	51.6

Table 1: Accuracy (%) in the *shitsukan* selection task, where models were prompted to choose the term appropriate for a specified object in a given image (§4). Models are sorted by their release date (newer on top). The strongest results in each category are **highlighted**, separately for open-source and proprietary models. *gpt-4-1106-vision-preview

of the prompts used in our evaluation is shown in Figure 16 in the appendix.

Measures. For the classification task, we used accuracy as the evaluation metric. In the generative setting, we calculated the overlap rate between generated and gold terms after normalizing the surface forms (i.e., lemmatization and conversion to *hiragana*). This measure, however, treats terms not recalled by humans as incorrect, which may not be the case. To address this, one of the authors, who is a native Japanese speaker, conducted a manual evaluation to assess the naturalness of the generated terms.[5] We compared classification in both Japanese and English, while the generation test was conducted in Japanese only.

4.4 Results

4.4.1 Classification.

The results are presented in Table 1. Models are sorted by recency (newest at the top) with proprietary models listed at the bottom, and this order appears to align with increasing performance scores. For instance, comparing the older LLaVA-1.5$_{7B}$ to the newer LLaVA-OneVision$_{7B}$ reveals a significant performance improvement (e.g., 52.5%→79.4% in the binary setting) despite sharing a similar architecture, likely due to advancements in learning strategy and training data.

However, open-source models still lag behind proprietary ones. The highest overall performance was achieved by GPT-4o in the binary setting in Japanese (93.7%), whereas the best-performing open-source model, LLaVA-OneVision$_{7B}$, scored 79.4% in the same setting. GPT-4o was also more robust to increasing the number of candidates from two to five, losing only 12.8 points compared to LLaVA-OneVision$_{7B}$'s 16.4-point drop.

The newest models, Qwen2-VL$_{7B}$, GPT-4V, and GPT-4o, consistently performed better in Japanese than in English (e.g., +3.6%, +10.7%, and +6.6% in five-way classification, respectively). In contrast, the other models showed poorer performance in Japanese, such as Llama-3.2-Vision$_{11B}$, which dropped by 26.6 points in binary classification.

4.4.2 Generation.

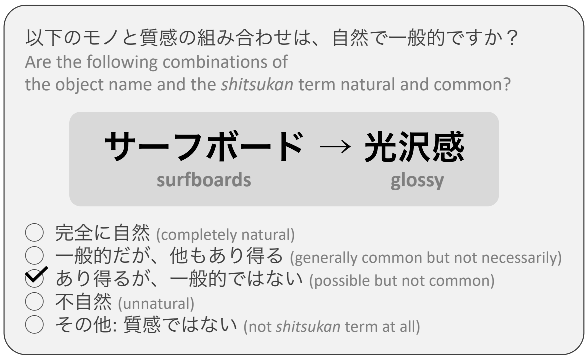
At the time when we conducted the manual annotation, the best-performing model in the classification task was GPT-4V[6] While its generated terms only had a 6.6% overlap rate with human-written terms, 268 out of 300 terms were still deemed natural in the manual evaluation, revealing a potential gap between *valid* terms and *commonly-recalled* terms.

One possible factor contributing to this gap could be a bias toward terms frequently appearing in the training data. To explore this, we analyzed the overlap between terms found in COCO captions, which are often included in LVLM training datasets, and those generated by the model. Specifically, we used Mecab (Kudo, 2005) to extract adjectives and adverbs from the Japanese captions of COCO images (STAIR Captions; Yoshikawa et al., 2017) and calculated their overlap with *shitsukan* terms generated by the best-performing open-source[7] LVLM in Japanese, Qwen2-VL$_{7B}$. Among 10,000 test cases (29,897 generated terms), only 17 overlapping terms were found (0.057%), indicating that the disparity between human- and model-generated terms is unlikely to be attributed to overlap with COCO captions.

[5] The annotator followed the same guidelines as those described in Section 5.1, with the addition of seeing the target image.

[6] gpt-4-1106-vision-preview, denoted as GPT-4V* in Table 1. Others (not shown) often did not successfully follow the instructed format.

[7] We excluded proprietary models from this analysis due to a lack of information on their training data.



以下のモノと質感の組み合わせは、自然で一般的ですか？
Are the following combinations of
the object name and the *shitsukan* term natural and common?

サーフボード → 光沢感
surfboards glossy

○ 完全に自然 (completely natural)
○ 一般的だが、他もあり得る (generally common but not necessarily)
✓ あり得るが、一般的ではない (possible but not common)
○ 不自然 (unnatural)
○ その他: 質感ではない (not *shitsukan* term at all)

Figure 5: Example of the task for evaluating the naturalness of *shitsukan* terms (§5.1). English translations are added for clarity and were not present in the original forms.

5 Commonsense Knowledge

Next, we examine *conceptual* understanding. Different from the task of producing *shitsukan* terms prompted by images, knowing plausible attributes of an object is a task touching on the *commonsense knowledge* of the model. Thus, here we ask: ***do models know which shitsukan terms are typical for an object?***

5.1 Task Design

In the naturalness evaluation task, models (and humans) are asked to evaluate how natural a *shitsukan* term is to describe a given object (Figure 5). We defined the following labels to categorize the appropriateness of the term:

- ***Completely natural***: typical for the object; it would be extremely rare for it not to apply.
- ***Typical, but not necessarily***: commonly applies, but alternatives are possible.
- ***Possible, but uncommon***: not typical, but it is not completely out of place.
- ***Unnatural***: not appropriate for the object.
- ***Other: not shitsukan***: not a *shitsukan* term at all. We used this option to flag invalid samples.

This somewhat fine-grained categorization is intended to contrast *shitsukan* terms that describe the object's *state* to what may be considered *typical* for it. For instance, while it may be typical for scissors to have a hard texture, a very worn, rusted, and old impression would be limited to well-used ones. Thus, we hypothesized that the former might be considered "typical" or "completely natural", and the latter as "possible, but uncommon."

We also conducted a generative variant of the task, where the task is to list three *shitsukan* terms that feel natural for a given object, *without* the context of an image. Similarly to the perceptual understanding task, we prepare these benchmarks by collecting additional human responses as described below.

5.2 Benchmark Construction

For the generative setting, we asked workers to list three *shitsukan* terms that they felt were "typical" for the specified object. Responses collected from three workers for each of the 169 object tags resulted in a collection of 1,443 terms after cleaning. As for the classification task, we asked workers to evaluate the appropriateness of a given *shitsukan* term and object pair as per our task definition. We collected judgments for 13,392 unique object-term pairs, each annotated by five workers. The gold labels were determined with a majority vote. Additional crowdsourcing details can be found in Appendix A.3.

5.3 Experimental Settings

Models. For classification, we compared the following models: Llama2-7b$_{\text{CH}}$ (Touvron et al., 2023), Llama2-7b$_{\text{JA-FI}}$ (Sasaki et al., 2023), GPT-3.5 (Brown et al., 2020), and GPT-4 (OpenAI, 2023).[8] For the generation task, we exclusively assessed GPT-4, as it was a challenging task that requires a high level of commonsense knowledge. The models were prompted in a few-shot manner. An example prompt is shown in Figure 17 in the appendix.

5.3.1 Measures.

For classification, we defined a measure expressing the extent to which the distribution of majority-vote labels aligns with the labels the model outputs. We first calculated the distribution of correct labels for each pair of object tags and list of terms (`label_list`): we calculated a normal distribution by computing the mean μ and standard deviation σ, where $i \in \{1, 2, \ldots, N\}$:

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

[8] Specifically, `Llama-2-7b-hf-chat`, `ELYZA-japanese-Llama-2-7b-fast-instruct`, `gpt-3.5-turbo-0301`, and `gpt-4-0314`, respectively.

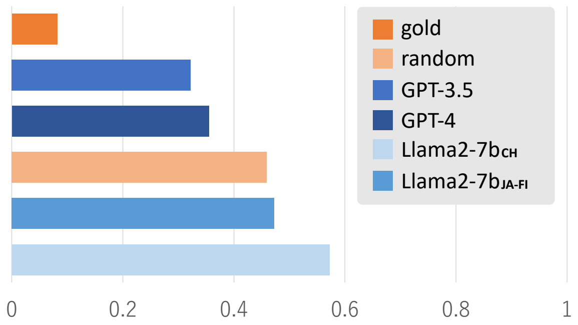


Figure 6: Results of the classification task for evaluating the naturalness of *shitsukan* terms (§5.1). A higher score indicates a greater disparity between the distribution of gold labels and the model's predicted labels.

Next, we calculated the distance d_i between the correct and predicted labels' distribution. The distance d_i is defined as the difference between the probability value of the most frequent label in the `label_list` $f_i(l_{gold})$ and the probability value of the predicted label $f_i(l_{pred})$:

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

Finally, we calculated the average distance value d_i to derive the overall score S $(0 < S < 1)$. A higher score of S indicates a greater disparity between the distribution of correct labels and the model's predicted labels.

For the generation task, we used the same measures as for the perceptual understanding tasks, i.e., overlap rate and a manual evaluation of 100 test samples for the naturalness of the generated terms.

5.4 Results

Classification. The results of the experiment are shown in Figure 6. Llama2-7b$_{\text{JA-FI}}$ and Llama2-7b$_{\text{CH}}$ produced a disparity equivalent to or larger than with random labels, suggesting a significant divergence from human judgments of *shitsukan* naturalness. In contrast, GPT-3.5 and GPT-4 had a lower disparity, indicating a similarity to human label distribution. However, a significant difference remains, suggesting that even the current SoTA model, GPT-4, does not fully emulate human *shitsukan* understanding.

Generation. The final score was 13.98% in the automated evaluation, suggesting a low likelihood of similarity between terms recalled by humans and those generated by GPT-4. However, 249 out of 300 generated terms were judged as natural, again demonstrating the pattern of GPT-4 generating different but natural terms.

6 Taxonomic Understanding

Finally, from a somewhat different angle, we consider general knowledge of *shitsukan* to include *taxonomic knowledge* of the concept. Japanese speakers are able to, with some variation, intuitively determine whether a word is a *shitsukan* term or not independently of its context. Here, we evaluate the models' ability to do the same. Our final question is then: ***do models generally recognize shitsukan terms?***

6.1 Task Design

Given our definition of *shitsukan*, the task is to recognize if a given word is a *shitsukan* term. We created several variants of this setting, depending on which and how many terms we use as negative candidates. In all settings, we used the terms labeled as *"completely natural"* and *"typical, but not necessarily"* in the naturalness assessment task (§5.1) as positive examples.

Yes/No binary classification. Answer whether a word is a *shitsukan* term or not. We created an easier and harder setting, depending on the source of negative examples: (1a) randomly selected adjectives and adjectival verbs from Wikipedia,[9] and (1b) terms flagged as non-*shitsukan* terms in the naturalness assessment task.

A/B binary classification. Using the same negative examples as in the Yes/No binary classification, we created a binary choice between a positive and negative candidate sample (2a and 2b, respectively). The task is to select the *shitsukan* term from the two choices.

Multiple-choice. Using the same negative examples as in (1b) and (2b), we created multiple-choice classification tasks with two to five candidates and one correct choice.

6.2 Experimental Settings

We compared the same models as in the previous experiments: Llama2-7b$_{\text{CH}}$, Llama2-7b$_{\text{JA-FI}}$, GPT-3.5, and GPT-4. For all settings, we measure the accuracy of the model's predictions. An example prompt is shown in Figure 18 in the appendix.

6.3 Results

Table 2 shows the results of Yes/No classification on the left side (1a, 1b), and A/B classification

[9] E.g., *dangerous*, *sinful*, etc. These are not typically *shitsukan* terms.

Lang	Model	Setting ID			
		1a	1b	2a	2b
JA	GPT-3.5	65.7	67.2	80.7	65.2
	GPT-4	74.3	72.0	81.8	79.0
	Llama2-7b$_{\text{CH}}$	50.9	50.9	50.9	51.2
	Llama2-7b$_{\text{JA-FI}}$	65.6	58.3	52.5	49.8
EN	GPT-3.5	65.5	51.1	69.0	50.8
	GPT-4	74.8	52.4	81.8	54.1
	Llama2-7b$_{\text{CH}}$	78.3	52.2	50.3	49.7

Table 2: Accuracy(%) for yes/no judgments (1.a, 1.b) and a/b selection (2.a, 2.b). Note that the chance rate is 50% for all settings.

Lang	Model	Number of Choices			
		2	3	4	5
JA	GPT-3.5	82.7	70.0	68.2	52.7
	GPT-4	87.3	67.3	67.3	60.9
	Llama2-7b$_{\text{CH}}$	57.3	34.6	26.4	26.4
	Llama2-7b$_{\text{JA-FI}}$	48.2	28.2	29.1	20.9
EN	GPT-3.5	50.0	40.0	24.9	23.6
	GPT-4	59.1	45.5	24.6	36.4
	Llama2-7b$_{\text{CH}}$	49.4	27.3	24.6	25.5

Table 3: Accuracy(%) for experiment 1.3. The chance rate for each row is 1 / {Number of Choices}.

on the right side (2a, 2b). Despite providing the definition of *shitsukan* terms, the Yes/No classification was solved with an accuracy of 78% or less, and the A/B binary classification with 82% or less. Several settings did not go above the chance rate despite additional pre-training in Japanese (e.g., Llama2-7b$_{\text{JA-FI}}$). GPT-4 seemed to perform best in both languages.

In multi-choice settings, shown in Table 3, performance decreased as the number of alternatives increased. Even GPT-4, the best-performing system in binary classification, dropped to an accuracy of 61% and 36% in Japanese and English, respectively. Overall, performance was consistently stronger in Japanese, suggesting that those models are better able to recall *shitsukan*-related knowledge given a Japanese context.

7 Discussion

We considered three aspects of understanding *shitsukan*: perception, commonsense knowledge, and taxonomic understanding. For perception, we found that LVLMs generated different but still natural *shitsukan* terms given an image and target object. Encouragingly, newer models improved over older variants, especially in Japanese. The latter is a surprising result considering that all models we used in our experiments are *English-centric* (pre-trained and fine-tuned primarily on English data), suggesting that models are more capable of eliciting *shitsukan*-related knowledge given a Japanese context. In turn, the lower performance in English warrants further research to understand whether these are technical limitations or due to cultural differences.

In contrast, LLMs showed low overlap with human distributions of assessments on the general naturalness of *shitsukan* terms for given objects. Generative results were similar to those with images, i.e., models generated different but still natural *shitsukan* terms given only a target object. As for taxonomic knowledge, models performed moderately on binary classification tasks, but degrading performance with a higher number of choices.

Overall, we found that there is evidence of *shitsukan* understanding in stronger models, particularly GPT-4. However, performance was near-chance in some models, so we conclude that there is significant room for improvement in LLM/LVLM alignment with human understanding of *shitsukan* for those models. On the other hand, strong generative performance hints towards potential strengths, i.e., models may be able to produce *shitsukan* terms with greater coverage and diversity than crowd-workers. This can be leveraged in future work to improve the quality of *shitsukan* datasets, or to provide a more comprehensive understanding of *shitsukan* terms in general.

8 Conclusion

This study evaluated the ability of LLMs/LVLMs to understand the concept of *shitsukan*. *Shitsukan* sits at the intersection of language and perception, representing a test bed for a fuzzily defined sensory experience as well as a non-Western concept. We crowdsourced a dataset of *shitsukan* terms and built benchmarks to test alignment with human-elicited terms and judgments. Overall, performance was mixed, with newer models improving over older variants. Generative settings resulted in low overlap with humans but were still natural to human observers, pointing towards possible future uses in generating *shitsukan* terms. Our dataset and benchmarks will contribute to future research to enable LLM/LVLM to acquire a sense of perception closer to that of humans.

Limitations

Scope. This study focused on perception, commonsense knowledge, and taxonomic understanding of *shitsukan*. However, other dimensions, such as *producing shitsukan* through visual modalities, remain unexplored. Recent advancements in multimodal models with image generation capabilities open possibilities for future research in this direction, including augmenting the *shitsukan* dataset. Furthermore, our evaluation centers on human alignment and does not address real-world applications, such as dialogue systems. Investigating how our findings apply to such use cases is left for future work. Additionally, while we ensured a fair comparison by providing the same information to humans and models, further steps—such as offering more examples, longer explanations, or even fine-tuning—could potentially improve model performance. However, improving models was beyond the scope of this study, and our analysis was kept straightforward. Finally, recent advances in neuron analysis could enable a more direct assessment of whether models encode *shitsukan*-related knowledge, which may also bypass the issue of potential performance loss due to imperfect instruction-following. We leave such investigations for future work.

Data quality. While extensive measures were taken to ensure data quality, crowdsourcing has inherent limitations, such as incomplete or inaccurate responses (Daniel et al., 2018). Although additional filtering reduced noise, there is still room for improving both data quality and coverage. Additionally, the English portion of our bilingual dataset is a translation of the Japanese side, rather than being independently collected. This may introduce biases and fail to capture the perspective of native English speakers in the same way as the original Japanese data. Lastly, while our experiments were designed to compare models and humans as directly as possible, collecting human performance on the automatically constructed benchmarks would be a valuable addition for future comparisons. Without it, our study assumes 100% accuracy as being the goal, while in reality, there might be room for legitimate subjective variance. Clarifying this point is left to future work.

Ethical Considerations

We used human crowdsourcing to collect *shitsukan* terms. During our collection, we ensured fair payment and provided clear instructions to workers. We also included a validation question to ensure the quality of the data. We anonymized the data and did not collect any personal information. To avoid harmful samples, we filtered out inappropriate responses and did not use source data samples that may contain people or disturbing content.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Numbers JP20H05952, JST Moonshot R&D Grant Number JPMJMS2011 (fundamental research), and the R&D Hub Aimed at Ensuring Transparency and Reliability of Generative AI Models project of the Ministry of Education, Culture, Sports, Science and Technology.

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.

Edward H Adelson. 2001. On seeing stuff: the perception of materials by humans and machines. In *Human Vision and Electronic Imaging VI*. SPIE.

Miriam Amin and Manuel Burghardt. 2020. A Survey on Approaches to Computational Humor Generation. In *Proceedings of the 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 29–41, Online. International Committee on Computational Linguistics.

BIG bench authors. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33:*

Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. 2014. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613.

Florian Daniel, Pavel Kucherbaev, Cinzia Cappiello, Boualem Benatallah, and Mohammad Allahbakhsh. 2018. Quality Control in Crowdsourcing: A Survey of Quality Attributes, Assessment Techniques, and Assurance Actions. *ACM Comput. Surv.*, 51(1):1–40.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Roland W Fleming, Karl R Gegenfurtner, and Shin'ya Nishida. 2015a. Visual perception of materials: the science of stuff. *Vision Res.*, 109:123–124.

Roland W Fleming, Shin'ya Nishida, and Karl R Gegenfurtner. 2015b. Perception of material properties. *Vision Res.*, 115(Pt B):157–162.

Fumiyo Hayakawa, Yukari Kazami, Katsuyoshi Nishinari, Kana Ioku, Sayuri Akuzawa, Yoshimasa Yamano, Yasumasa Baba, and Kaoru Kohyama. 2013. Classification of Japanese texture terms. *J. Texture Stud.*, 44(2):140–159.

Jack Hessel, Ana Marasović, Jena D. Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. 2023. Do androids laugh at electric sheep? humor "understanding" benchmarks from the new yorker caption contest. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 688–714. Association for Computational Linguistics.

Hidehiko Komatsu and Naokazu Goda. 2018. Neural Mechanisms of Material Perception: Quest on Shitsukan. *Neuroscience*, 392:329–347.

Taku Kudo. 2005. Mecab: Yet another part-of-speech and morphological analyzer.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. 2024. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *Computer Vision – ECCV 2014*, pages 740–755. Springer International Publishing.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024b. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning. *Adv. Neural Inf. Process. Syst.*

Laurence T Maloney and David H Brainard. 2010. Color and material perception: Achievements and challenges. *J. Vis.*, 10(9).

Phillip J Marlow, Juno Kim, and Barton L Anderson. 2012. The perception and misperception of specular surface reflectance. *Curr. Biol.*, 22(20):1909–1913.

Daphne Maurer, Thanujeni Pathman, and Catherine J Mondloch. 2006. The shape of boubas: sound-shape correspondences in toddlers and adults. *Dev. Sci.*, 9(3):316–322.

Isamu Motoyoshi, Shin'ya Nishida, Lavanya Sharan, and Edward H Adelson. 2007. Image statistics and the perception of surface qualities. *Nature*, 447(7141):206–209.

S Nishida and M Shinya. 1998. Use of image-based information in judgments of surface-reflectance properties. *J. Opt. Soc. Am. A Opt. Image Sci. Vis.*, 15(12):2951–2965.

Katsuyoshi Nishinari, Fumiyo Hayakawa, Chong-Fei Xia, Long Huang, Jean-François Meullenet, and Jean-Marc Sieffermann. 2008. Comparative study of texture terms: English, French, Japanese and Chinese. *J. Texture Stud.*, 39(5):530–568.

Shogo Okamoto, Hikaru Nagano, and Yoji Yamada. 2013. Psychophysical dimensions of tactile perception of textures. *IEEE Trans. Haptics*, 6(1):81–93.

OpenAI. 2023. GPT-4 Technical Report. *Preprint*, arXiv:2303.08774.

Cory Paik, Stéphane Aroca-Ouellette, Alessandro Roncone, and Katharina Kann. 2021. The world of an octopus: How reporting bias influences a language model's perception of color. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 823–835. Association for Computational Linguistics.

Maki Sakamoto and Junji Watanabe. 2018. Bouba/Kiki in Touch: Associations Between Tactile Perceptual Qualities and Japanese Phonemes. *Front. Psychol.*, 9:295.

Akira Sasaki, Masato Hirakawa, Shintaro Horie, and Tomoaki Nakamura. 2023. ELYZA-japanese-Llama-2-7b.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-OKVQA: A Benchmark for Visual Question Answering Using World Knowledge. In *Computer Vision – ECCV 2022*, pages 146–162. Springer Nature Switzerland.

Izuru Shinmura. 2018. *Kojien*. Iwanami Shoten Tokyo.

Charles Spence. 2020. Shitsukan - the Multisensory Perception of Quality. *Multisens Res*, 33(7):737–775.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4149–4158. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. *CoRR*, abs/2302.13971.

Atsushi Wada, Yuichi Sakano, and Hiroshi Ando. 2014. Human cortical areas involved in perception of surface glossiness. *Neuroimage*, 98:243–257.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Bodo Winter, Márton Sóskuthy, Marcus Perlman, and Mark Dingemanse. 2022. Trilled /r/ is associated with roughness, linking sound and touch across spoken languages. *Sci. Rep.*, 12(1):1035.

Chenyun Wu, Mikayla Timm, and Subhransu Maji. 2020. Describing textures using natural language. In *European Conference on Computer Vision*, pages 52–70. Springer.

Shuquan Ye, Yujia Xie, Dongdong Chen, Yichong Xu, Lu Yuan, Chenguang Zhu, and Jing Liao. 2023. Improving commonsense in vision-language models via knowledge graph riddles. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Yuya Yoshikawa, Yutaro Shigeto, and Takeuchi Akikazu. 2017. STAIR captions: Constructing a large-scale japanese image caption dataset. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 417–421, Vancouver, Canada. Association for Computational Linguistics.

A Crowdsourcing Details

A.1 Main Data Collection

Qualifications. All workers were chosen from registered users on the Yahoo! Crowdsourcing platform who were native Japanese speakers over the age of 18. Their identities remained anonymous during the collection process, with just their IDs being made publicly available.

To select suitable workers, we designed a pilot round which assessed whether they could understand the task guidelines and provide appropriate answers. We hand-crafted a set of twelve questions with our own gold answers, and only kept workers who achieved an acceptable score. 200 workers participated initially, with and additional 108 participants in a later second call (308 total). After every two rounds, we also manually examined 20 random responses from each worker and manually tagged those that should be excluded. Grounds for exclusion included gibberish responses, non-conforming to the task format, and terms that are clearly not *shitsukan*. Ultimately, our whitelist was reduced to 60 workers after the final round of manual filtering, out of 131 initially qualified workers.

Compensation. The Yahoo! Crowdsourcing platform only allows a point-based compensation system, where each point is worth 1 JPY. For each task type, the platform sets a minimum point compensation and requesters are free to add additional points. The fee differs depending on the number of points added. For simplicity, we will report the total compensation in points and the final costs only (§A.4).

Qualification rounds, with three samples per HIT, were compensated with 45 points per HIT (roughly $0.30). For the main task, workers were given 150 points per HIT with 10 samples (with the exception of the last round where each HIT had 8 samples, i.e., 120 points). This is equivalent to roughly $1.00 per HIT, which matches the minimum wage in Japan considering the required time. With a total of 1,016 HITs divided in 11 rounds and an average of 96 qualified workers (131 workers in the first round and 60 in the last), a single worker could expect to earn around 1,570 points (around $10.00) in total.

Forms. A screenshot of the instruction screen and collection form used in the main data collection is shown in Figure 8 and Figure 10.

A.2 Task A: *Shitsukan* Term Selection

Qualifications. We manually prepared a set of 30 questions and only kept workers with a high score. Out of 600 participants, 48 passed our qualification test.

Compensation. In both qualification and main rounds, workers were compensated 20 points per HIT containing 10 samples. This amounts to 20 JPY (roughly $0.13) per HIT, matching the minimum wage in Japan considering the required time. With a total of 13,325 HITs available to a single worker and 48 qualified workers, she could expect to earn around 5,510 points (around $35.00) in total.

Preprocessing. The forms were prepared by sorting all terms into triples (positive candidates) associated with the target image and object, then adding two random negative candidates to each. As a result, most terms appear at least once as a candidate.

We prepared forms for 2,697 target image and object pairs and all their associated candidate terms, so that each appears at least once as a candidate (5,321 HITs). In total, Each was annotated by 5 workers.

Forms A screenshot of the instruction screen and collection form is shown in Figure 9 and Figure 11. Translations are available upon request to the authors.

Data example. Figure 12 shows an example of the voting data collected in the *shitsukan* selection task.

A.3 Task B: *Shitsukan* Term Naturalness Evaluation

Qualifications. Similarly to the selection task (§A.2), we prepared a set of 30 questions and a range of acceptable answers which we used to filter workers. Out of 600 participants, 46 passed the qualification test.

Compensation. Workers were compensated 20 points per HIT with 10 samples. This amounts to 20 JPY (roughly $0.13) per HIT, matching the minimum wage in Japan considering the required time. With a total of 6,700 HITs and 46 qualified workers, a single worker could expect to earn around 2,913 points (around $18.50).

Forms A screenshot of the instruction screen and collection form is shown in Figure 14 and Figure 15.

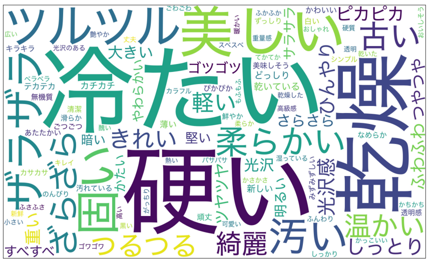


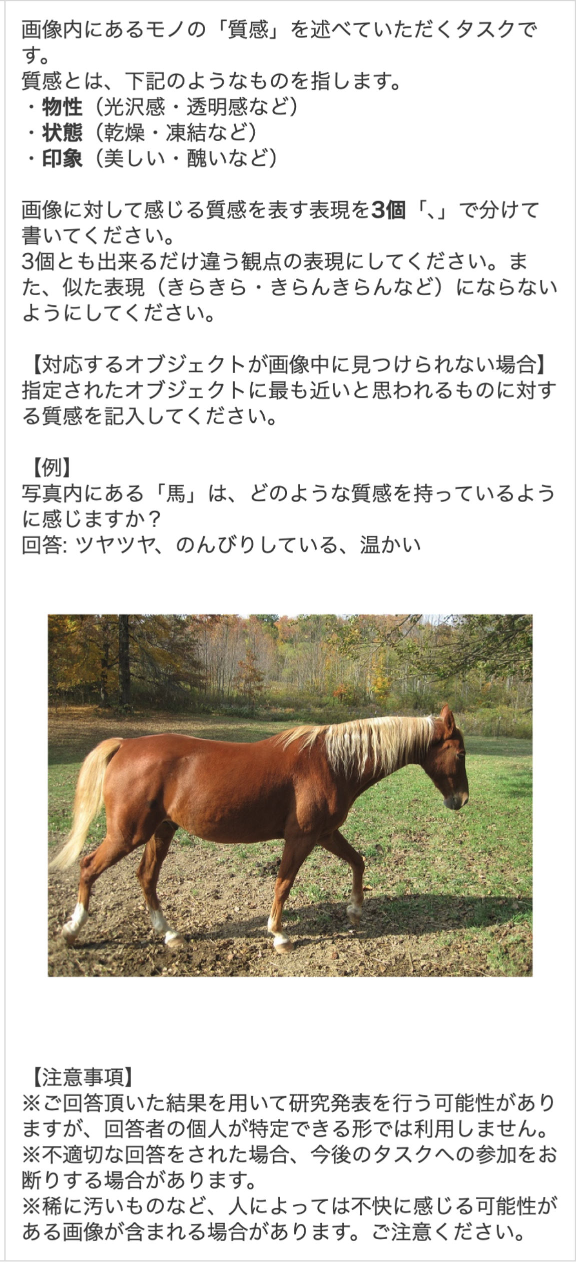
Figure 7: Word cloud of the most frequent Japanese *shitsukan* terms in the dataset.

Translations are available upon request to the authors.

Data example. Figure 13 shows an example of the data collected in the *shitsukan* term appropriateness assessment task.

A.4 Total Cost

The main data collection, including platform fees, cost 243,125 JPY. Task A cost an additional 99,671 JPY, and Task B cost 250,080 JPY. The total cost of all qualification rounds was 80,876 JPY, and we spent an additional 14,778 JPY on various trials. In total, this amounts to 688,530 JPY, or approximately $4,380 at the time of writing.



画像内にあるモノの「質感」を述べていただくタスクです。
質感とは、下記のようなものを指します。
・**物性**（光沢感・透明感など）
・**状態**（乾燥・凍結など）
・**印象**（美しい・醜いなど）

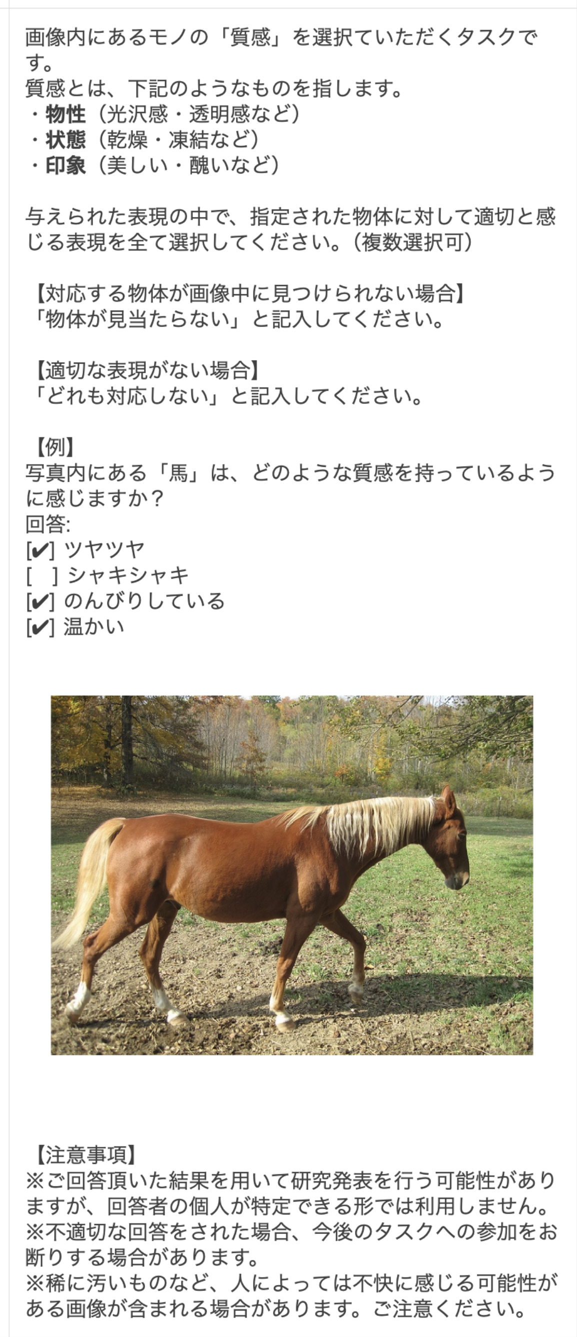
画像に対して感じる質感を表す表現を**3個**「、」で分けて書いてください。
3個とも出来るだけ違う観点の表現にしてください。また、似た表現（きらきら・きらんきらんなど）にならないようにしてください。

【対応するオブジェクトが画像中に見つけられない場合】
指定されたオブジェクトに最も近いと思われるものに対する質感を記入してください。

【例】
写真内にある「馬」は、どのような質感を持っているように感じますか？
回答: ツヤツヤ、のんびりしている、温かい

【注意事項】
※ご回答頂いた結果を用いて研究発表を行う可能性がありますが、回答者の個人が特定できる形では利用しません。
※不適切な回答をされた場合、今後のタスクへの参加をお断りする場合があります。
※稀に汚いものなど、人によっては不快に感じる可能性がある画像が含まれる場合があります。ご注意ください。

Figure 8: Instructions shown to workers during for the main *shitsukan* data collection (§3).



画像内にあるモノの「質感」を選択ていただくタスクです。
質感とは、下記のようなものを指します。
・**物性**（光沢感・透明感など）
・**状態**（乾燥・凍結など）
・**印象**（美しい・醜いなど）

与えられた表現の中で、指定された物体に対して適切と感じる表現を全て選択してください。（複数選択可）

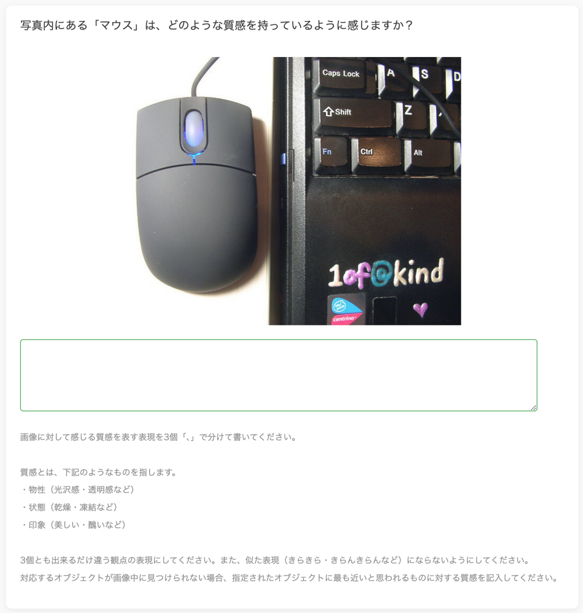
【対応する物体が画像中に見つけられない場合】
「物体が見当たらない」と記入してください。

【適切な表現がない場合】
「どれも対応しない」と記入してください。

【例】
写真内にある「馬」は、どのような質感を持っているように感じますか？
回答:
[✔] ツヤツヤ
[　] シャキシャキ
[✔] のんびりしている
[✔] 温かい

【注意事項】
※ご回答頂いた結果を用いて研究発表を行う可能性がありますが、回答者の個人が特定できる形では利用しません。
※不適切な回答をされた場合、今後のタスクへの参加をお断りする場合があります。
※稀に汚いものなど、人によっては不快に感じる可能性がある画像が含まれる場合があります。ご注意ください。

Figure 9: Instructions given to workers for the *shitsukan* selection task (§4.1).



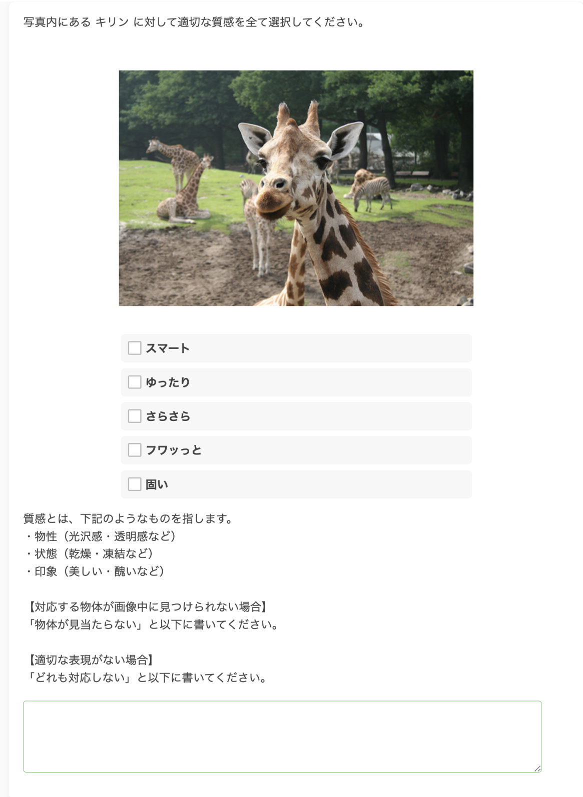
写真内にある「マウス」は、どのような質感を持っているように感じますか？

画像に対して感じる質感を表す表現を3個「、」で分けて書いてください。

質感とは、下記のようなものを指します。
・物性（光沢感・透明感など）
・状態（乾燥・凍結など）
・印象（美しい・醜いなど）

3個とも出来るだけ違う観点の表現にしてください。また、似た表現（きらきら・きらんきらんなど）にならないようにしてください。
対応するオブジェクトが画像中に見つけられない場合、指定されたオブジェクトに最も近いと思われるものに対する質感を記入してください。

Figure 10: Example of a form used in the main data collection (§3).



写真内にある キリン に対して適切な質感を全て選択してください。

□ スマート
□ ゆったり
□ さらさら
□ フワッっと
□ 固い

質感とは、下記のようなものを指します。
・物性（光沢感・透明感など）
・状態（乾燥・凍結など）
・印象（美しい・醜いなど）

【対応する物体が画像中に見つけられない場合】
「物体が見当たらない」と以下に書いてください。

【適切な表現がない場合】
「どれも対応しない」と以下に書いてください。

Figure 11: Example of a form used in the *shitsukan* selection task (§4.1).

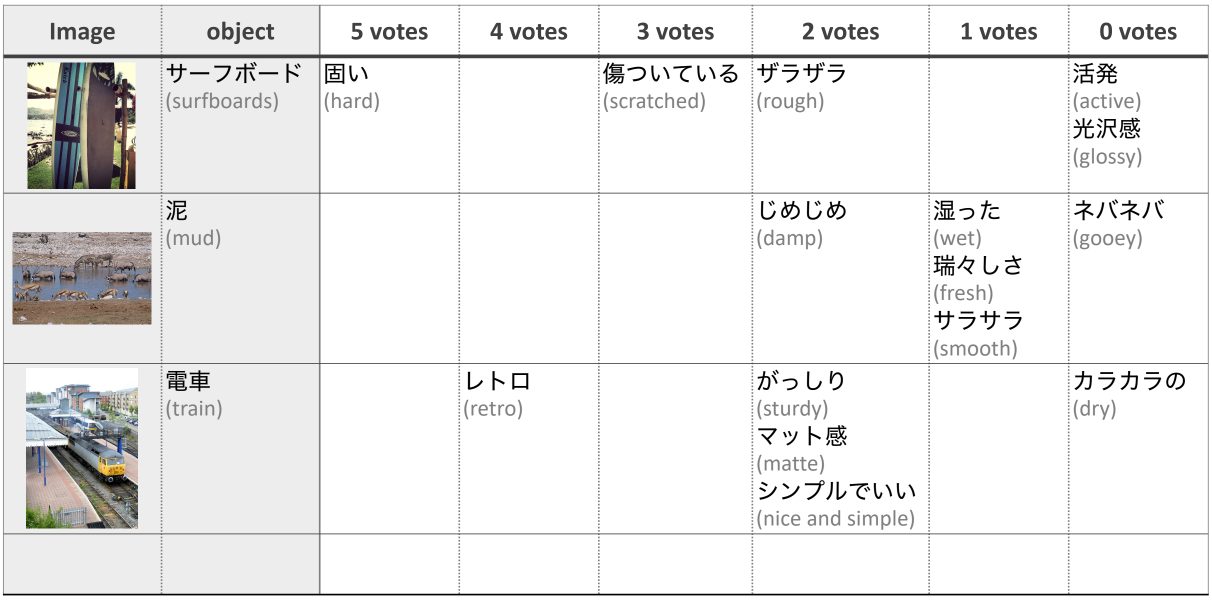
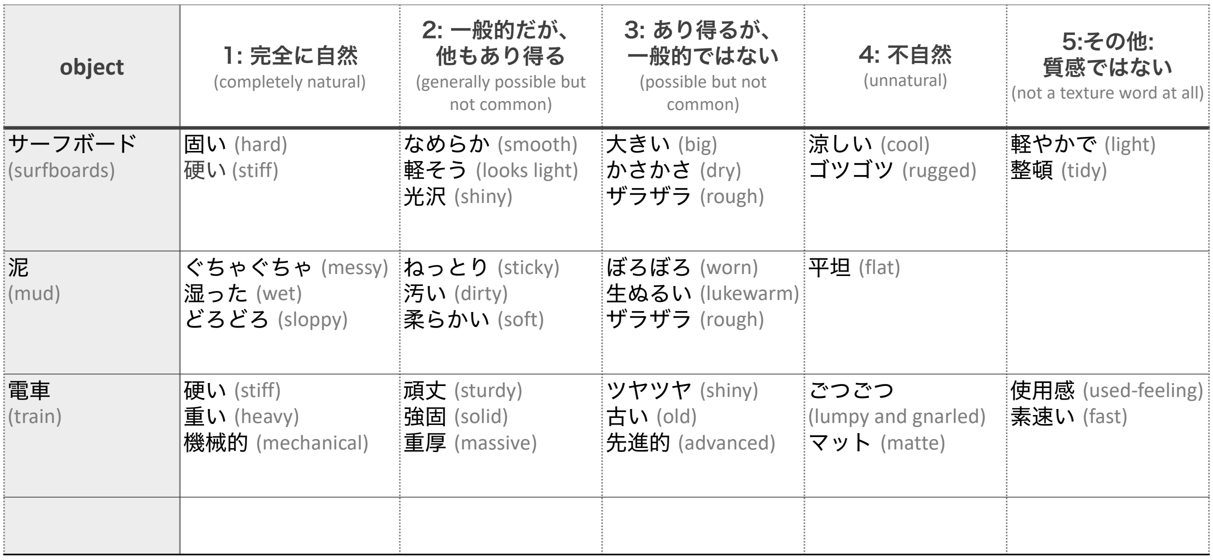


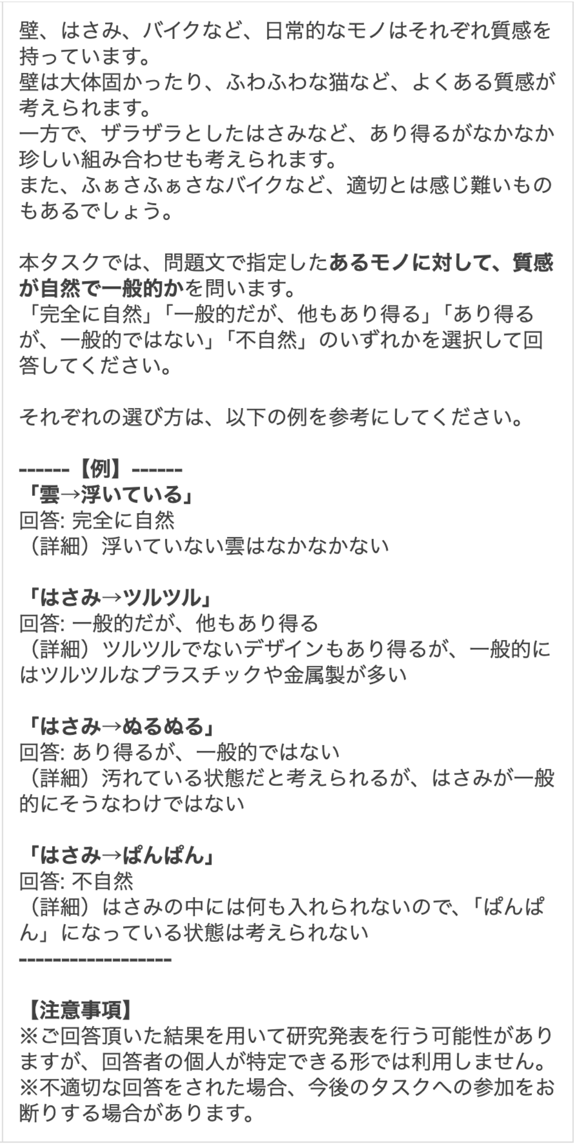
Image	object	5 votes	4 votes	3 votes	2 votes	1 votes	0 votes
	サーフボード (surfboards)	固い (hard)		傷ついている (scratched)	ザラザラ (rough)		活発 (active) 光沢感 (glossy)
	泥 (mud)				じめじめ (damp)	湿った (wet) 瑞々しさ (fresh) サラサラ (smooth)	ネバネバ (gooey)
	電車 (train)		レトロ (retro)		がっしり (sturdy) マット感 (matte) シンプルでいい (nice and simple)		カラカラの (dry)

Figure 12: Example of the voting data collected in the *shitsukan* selection task.



object	1: 完全に自然 (completely natural)	2: 一般的だが、他もあり得る (generally possible but not common)	3: あり得るが、一般的ではない (possible but not common)	4: 不自然 (unnatural)	5:その他: 質感ではない (not a texture word 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 (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)
泥 (mud)	ぐちゃぐちゃ (messy) 湿った (wet) どろどろ (sloppy)	ねっとり (sticky) 汚い (dirty) 柔らかい (soft)	ぼろぼろ (worn) 生ぬるい (lukewarm) ザラザラ (rough)	平坦 (flat)	
電車 (train)	硬い (stiff) 重い (heavy) 機械的 (mechanical)	頑丈 (sturdy) 強固 (solid) 重厚 (massive)	ツヤツヤ (shiny) 古い (old) 先進的 (advanced)	ごつごつ (lumpy and gnarled) マット (matte)	使用感 (used-feeling) 素速い (fast)

Figure 13: Example of the data collected in the *shitsukan* term appropriateness assessment task.



壁、はさみ、バイクなど、日常的なモノはそれぞれ質感を持っています。
壁は大体固かったり、ふわふわな猫など、よくある質感が考えられます。
一方で、ザラザラとしたはさみなど、あり得るがなかなか珍しい組み合わせも考えられます。
また、ふぁさふぁさなバイクなど、適切とは感じ難いものもあるでしょう。

本タスクでは、問題文で指定した**あるモノに対して、質感が自然で一般的か**を問います。
「完全に自然」「一般的だが、他もあり得る」「あり得るが、一般的ではない」「不自然」のいずれかを選択して回答してください。

それぞれの選び方は、以下の例を参考にしてください。

------【例】------
「雲→浮いている」
回答: 完全に自然
（詳細）浮いていない雲はなかなかない

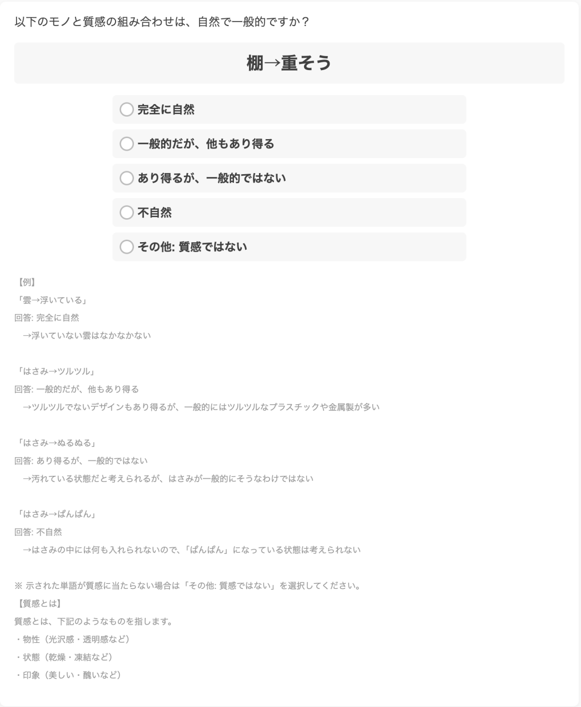
「はさみ→ツルツル」
回答: 一般的だが、他もあり得る
（詳細）ツルツルでないデザインもあり得るが、一般的にはツルツルなプラスチックや金属製が多い

「はさみ→ぬるぬる」
回答: あり得るが、一般的ではない
（詳細）汚れている状態だと考えられるが、はさみが一般的にそうなわけではない

「はさみ→ぱんぱん」
回答: 不自然
（詳細）はさみの中には何も入れられないので、「ぱんぱん」になっている状態は考えられない

【注意事項】
※ご回答頂いた結果を用いて研究発表を行う可能性がありますが、回答者の個人が特定できる形では利用しません。
※不適切な回答をされた場合、今後のタスクへの参加をお断りする場合があります。

Figure 14: Instructions given to workers for the naturalness evaluation task (§5.1).



以下のモノと質感の組み合わせは、自然で一般的ですか？

棚→重そう

○ 完全に自然
○ 一般的だが、他もあり得る
○ あり得るが、一般的ではない
○ 不自然
○ その他: 質感ではない

【例】
「雲→浮いている」
回答: 完全に自然
　→浮いていない雲はなかなかない

「はさみ→ツルツル」
回答: 一般的だが、他もあり得る
　→ツルツルでないデザインもあり得るが、一般的にはツルツルなプラスチックや金属製が多い

「はさみ→ぬるぬる」
回答: あり得るが、一般的ではない
　→汚れている状態だと考えられるが、はさみが一般的にそうなわけではない

「はさみ→ぱんぱん」
回答: 不自然
　→はさみの中には何も入れられないので、「ぱんぱん」になっている状態は考えられない

※ 示された単語が質感に当たらない場合は「その他: 質感ではない」を選択してください。
【質感とは】
質感とは、下記のようなものを指します。
・物性（光沢感・透明感など）
・状態（乾燥・凍結など）
・印象（美しい・醜いなど）

Figure 15: Example of a form used in the naturalness evaluation task (§5.1).

The task is to write terms that describe the “essence” of an object. The essence of an object refers to the following:

- Physical properties (glossiness, transparency, etc.)
- State (dry, frozen, etc.)
- Impression (beautiful, ugly, etc.)

Write three terms that express the essence of the specified object in the image separated by “, ”. Please try to make all three expressions as different as possible. Also, try to avoid using similar expressions (sparkle/sparkly, etc.).

[If you cannot find the specified object in the image]
Please write the essence of the object that you think is closest to the specified object.

[Example]
What kind of *shitsukan* do you feel the “horse” in the photo has?
Answer: shiny, relaxed, warm

What kind of essence do you think the “bowl” in the photo has?
Answer:

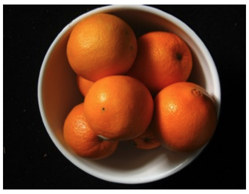


Figure 16: Prompt for the *shitsukan* description task to assess the perception (Section 4.3). English translation; the original is in Japanese. The image input is on the right. Note that some open-source models did not support multiple input images, so the image associated with the example is missing.

This task asks you to describe the “essence” a specified object generally has. The essence of an object refers to the following:

Shitsukan refers to the following:

- Physical properties (gloss, transparency, etc.)
- State (dry, frozen, etc.)
- Impression (beautiful, ugly, etc.)

Write three expressions that describe the essence you feel the specified object generally has, separated by “, ”. Please try to make all three expressions as different as possible. Also, try not to use similar expressions (sparkle/sparkly, etc.).

[Example]
What kind of essence do you think a “horse” generally has?
Answer: shiny, relaxed, warm

What kind of essence do you think a “carrot” generally has?
Answer:

Figure 17: *Shitsukan* description task prompt (Section 5.1). Translated to English for clarity; the original is in Japanese.

The task is to judge which of the given terms describe an “essence” of an object.
The essence of an object refers to the following:

- Physical properties (gloss, transparency, etc.)
- Condition (dry, frozen, etc.)
- Impression (beautiful, ugly, etc.)

Please select the term that describes an essence of an object.

[Example]
0: friendly, 1: stopped, 2: foreign, 3: matte
Answer: 3

Please select the term that describes an essence of an object.
0: shiny, 1: dormant, 2: atlantic, 3: human
Answer:

Figure 18: Example *shitsukan* taxonomic understanding task prompt (multiple-choice; §6).