RELATIVE DRAWING IDENTIFICATION COMPLEXITY IS INVARIANT TO MODALITY IN VISION-LANGUAGE MODELS*

Diogo Freitas
Interactive Technologies Institute
and NOVA LINCS
Faculty of Exact Sciences and Engineering
University of Madeira
Portugal
diogo.freitas@staff.uma.pt

Brigt Håvardstun
Department of Informatics
University of Bergen
Norway
brigt.havardstun@uib.no

Cèsar Ferri
Valencian Research Institute for
Artificial Intelligence
Universitat Politècnica de València
Spain
cferri@dsic.upv.es

Darío Garigliotti
Department of Informatics
University of Bergen
Norway
Dario.Garigliotti@uib.no

Jan Arne Telle
Department of Informatics
University of Bergen
Norway
Jan.Arne.Telle@uib.no

José Hernández-Orallo
Leverhulme Centre for the Future
of Intelligence and
Valencian Research Institute for
Artificial Intelligence
Spain
jorallo@upv.es

August 29, 2025

ABSTRACT

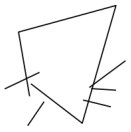
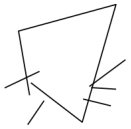
Large language models have become multimodal, and many of them are said to integrate their modalities using common representations. If this were true, a drawing of a car as an image, for instance, should map to a similar area in the latent space as a textual description of the strokes that form the drawing. To explore this in a black-box access regime to these models, we propose the use of machine teaching, a theory that studies the minimal set of examples a teacher needs to choose so that the learner captures the concept. In this paper, we evaluate the complexity of teaching vision-language models a subset of objects in the Quick, Draw! dataset using two presentations: raw images as bitmaps and trace coordinates in TikZ format. The results indicate that image-based representations generally require fewer segments and achieve higher accuracy than coordinate-based representations. But, surprisingly, the teaching size usually ranks concepts similarly across both modalities, even when controlling for (a human proxy of) concept priors, suggesting that the simplicity of concepts may be an inherent property that transcends modality representations.

1 Introduction

As children, when we transform images of the world into drawings and other simplified sketches, we have the intuition that some objects are simpler than others [5, 18]. For instance, six segments are enough to represent a `house` that everybody can recognize, while a bit more is necessary to represent a `cat`. This intuition is epitomized by some guessing games where one person picks a concept from a card deck and has to draw something quickly for their team to identify the concept. We can easily describe and recognize some very simple visual concepts, such as letters, with verbalized descriptions. For instance, the letter `T` is a horizontal segment on top of a vertical segment. However, humans struggle to describe more complex shapes with verbal descriptions [26] or objects, such as a `cat`, using a series of segments.

*This research was accepted at the 28th European Conference on Artificial Intelligence (ECAI-2025). Paper ID: 7633.

Table 1: The simplest drawings (applying RDP algorithm on an original drawing) identified for the concept `cat`.

Model	Original (images)	Simplified (images)	Original (coordinates)	Simplified (coordinates)
Claude				
Gemini				
GPT-4 Turbo				
GPT-4o				
Llama				
Pixtral				

Large Language Models (LLMs) can identify objects from a textual representation of their coordinates [3]. Thus, we aim to discover whether this understanding maps to similar capabilities for the multimodal versions of these models. Also, we do not know whether this is independent of the modality. We ask two research questions:

- Q1 (*Absolute Invariance*): If we randomly sample a concept from a concept class, $c \in C$, would it take the same number of segments to identify it if represented as a bitmap drawing as if represented as a set of coordinates?
- Q2 (*Relative Invariance*): If we randomly sample two concepts from a concept class, $c_1, c_2 \in C$, and c_1 requires fewer segments than c_2 when represented as a bitmap, will this order prevail when expressed as coordinates?

Question Q1 refers to whether a concept represented as a bitmap drawing is easier or harder to recognize than the same concept as coordinates in text, while question Q2 is about the relative ranking. For instance, consider that c_1 is a `house` and c_2 is a `cat`. In Figure 1, if a `house` is easier than a `cat` when using the bitmap of the drawing (top of the figure), is it also easier when represented as segment coordinates (bottom of the figure)? This is the *relative invariance*. Note that we are not comparing with photographic images of the object since other features would come into play, such as a striped texture to distinguish a `tiger` from other felines. Such distinctions are particularly evident in machine vision systems [11].

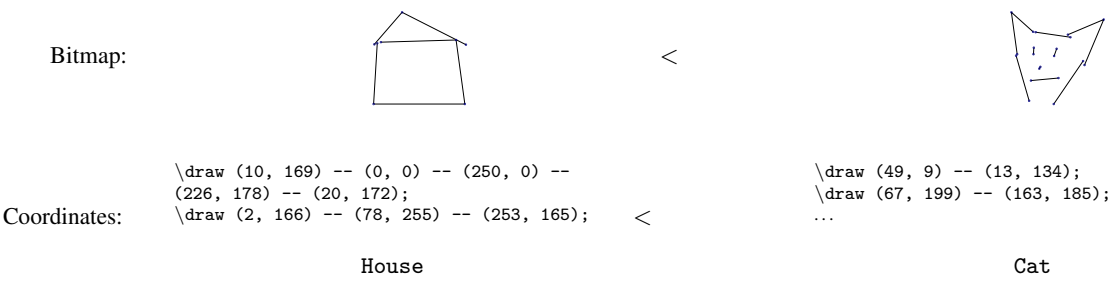


Figure 1: In this paper, we address two research questions. First, Q1 (absolute invariance): When using a vision-language model, are bitmaps (top) equally efficient representations for drawings as coordinates (bottom)? The second question is Q2 (relative invariance): Is the order (left vs. right) of simplicity preserved across modalities?

However, how can one determine the notion of simplicity of a concept from its drawings? The idea we pursue in this paper is based on the field of machine teaching [36], and in particular, the notion of teaching minimality. A concept is as simple as a teacher can communicate the concept to a learner with as little information as possible. This captures our intuition that a `house` needs six segments while a `cat` needs more segments. Given a concept, the teacher has to find the simplest drawing in terms of the number of straight-line segments—the teaching size—that enables the learner to consistently recognize the concept. We use two different types of language representations (bitmaps of the drawing and coordinates in TikZ code) to present the concepts to the learner. Multiple models, including Generative Pretrained Transformer (GPT)-4 [1], Llama [13], Gemini [29], Pixtral, and Claude, are employed as the "learners". The resulting collection of the simplest images identified, across all concepts, all modalities, and all models, is intriguingly diverse. As a preview of our findings, see Table 1, showing the simplest identified images for the concept `cat`.

It is also important to note that priors play a role in machine teaching. When in doubt, the learner will more likely associate the evidence with the most common concept (e.g., a `house` is more common than an `envelope`). Accordingly, a Bayesian prior will be used to disentangle this effect when looking at the concept simplicity rankings.

The contributions of this paper are:

- A novel machine teaching framework for evaluating the complexity of concepts, which can be applied to drawings in coordinate- and image-based modalities.
- Use of the teaching size specifically to evaluate how simply and effectively the concept can be taught across both modalities.
- A comparison of both modalities across multiple models, including GPT-4, Llama, Gemini, Pixtral, and Claude, according to the number of concepts identified, accuracy, frequency of errors, and teaching size.
- A way to disentangle the effect of the learner's prior knowledge in the concept identification task.

These contributions are generic and can be applied to other problems and modalities. In our particular case, we show that bitmaps are more efficient than coordinates, but surprisingly, the order of complexity between the concepts is preserved to some extent. This suggests that either the representations of both modalities are tightly connected in the latent space of the model, or the simplicity of concepts is an inherent property that transcends modalities.

2 Related Work

Drawing (or Sketches) Recognition: Eitz et al. [8] provided a dataset of human drawings, including 250 concepts and 20,000 drawings. They introduced a support vector machine model to recognize these drawings and observed that humans outperformed its performance. Since then, AI models have been closer or even achieved higher accuracy than that of human classification for drawing recognition (e.g., Schneider and Tuytelaars 24, Yu et al. 34, Zhang et al. 35, Yang et al. 33). Using the *Quick, Draw!* dataset, Ha and Eck [14] proposed `sketch-rnn`, a model designed to create drawings of common objects that resemble those drawn by humans. A similar version of this model has also shown capabilities in drawing recognition [2]. Other neural approaches studied for this task include convolutional neural networks [16], and graph neural networks applied over drawings represented as graphs [32].

Drawing Capacities of LLMs: Sharma et al. [25] assess the visual abilities of different language models. They conduct experiments that prompt the models to create code that draws images based on text descriptions and improve image generation code iteratively through text feedback. They show that: (a) LLMs possess limited ability to recognize concepts represented in code, and (b) these models sometimes fail to recognize concepts that they can accurately draw.

Note that the authors addressed the problem as a multi-class classification problem. Moreover, the online interface for collecting human drawings limits components to basic shapes like ellipses, possibly restricting participants' ability to create complex drawings. In their initial experiments with GPT-4, Bubeck et al. [3] present an example of drawing generation, showcasing text-to-image capabilities using TikZ. They show tasks such as GPT-4 drawing a unicorn and constructing TikZ code through a multi-step prompt process. In another study, Pourreza et al. [21] introduce the *Painter*, a modified LLM that creates drawings using virtual brush strokes based on user-provided text descriptions. Additionally, Cai et al. [4] evaluated GPT-4's ability to understand visual data in SVG format across various visual tasks, including image classification, visual reasoning, and image generation. Vinker et al. [31] propose *SketchAgent*, showing that while LLMs iteratively generate sketches, they struggle with spatial reasoning.

Machine Teaching: Machine teaching is a research area that focuses on identifying the optimal set of examples that allow a learner (e.g., a human or a machine) to identify a given concept [36]. To illustrate the underlying idea of machine teaching, assume the teacher wants the learner to identify the concept of prime numbers. To achieve this, the teacher uses the set $S_1 = \{2, 3, 5, 7, 11, 13\}$ and succeeds. However, would it not be enough for the learner just to see the smaller set $S_2 = \{19, 23\}$? Of course, that depends on the learner. In general, optimal teaching will depend on the model the teacher has of the learner. Machine teaching presents an alternative framework to machine learning (where examples are not chosen but sampled from a distribution) to answer the question of whether some concepts are inherently more complex than others. The connections between machine teaching and computational learning theory are strong; see, e.g., the works by Doliwa et al. [6] or Moran and Yehudayoff [19], with machine teaching putting the emphasis on the minimal evidence that distinguishes the concept from all the rest. To determine how easy it is to teach a concept, the teaching dimension [36]—the minimum number of examples the learner needs to identify a concept—was traditionally used. Telle et al. [30] introduced a new metric named teaching size. This metric puts the focus on the sum of the sizes of the examples needed to identify a concept, rather than only the number of examples.

3 Methods

The drawings used in this work come from the *Quick, Draw!* dataset [15, 14], which includes over 50 million drawings of 345 concepts. Collected by Google Creative Lab via an interactive game, participants had 20 seconds to draw a concept while a neural network attempted real-time recognition. The dataset is the largest collection of doodles in the world, with contributions from more than 15 million participants.

Each drawing in the Simplified Drawing files that we use is stored as vectors of distinct pen strokes, i.e., distinct continuous movements of the pen without lifting. Each stroke s_i is represented by a sequence of (x, y) coordinates $\{(x_{i1}, y_{i1}), (x_{i2}, y_{i2}), \ldots, (x_{in}, y_{in})\}$. Note that each pair of consecutive points in a stroke creates a segment. Additionally, for each drawing, a binary flag r indicates whether the game's neural network correctly recognized the concept.

The following sections cover concept selection, corresponding drawings, learners, the machine teaching setting, and the drawing selection conducted before testing the framework.

3.1 Teaching Size

Let D denote an infinite space of possible drawings (and their simplifications, as will be explained later), and let C be a set of concepts. We use D_c to denote all the drawings of a concept $c \in C$. For any given concept $c \in C$, the objective is to identify the simplest drawing $S \in D_c$ (represented as S^m with modality m being either bitmap or coordinates) such that a learner L successfully learns c with a probability of at least ρ over N independent trials (i.e., recognition consistency). The *teaching size (TS)* of c for the modality m can then be defined as follows:

$$\mathrm{TS}_{\rho,N,m}(c) = \min_{S \in D_c} |S^m| \text{ s.t. } \sum_{1}^{N} \mathbb{1}\left[L(S^m) = c\right] \geq \rho \cdot N, \tag{1}$$

where $\mathbb{1}[\cdot]$ is the indicator function, which equals 1 if the learner L correctly identifies concept c from the drawing S^m, and 0 otherwise.

We argue that a good metric for assessing the simplicity of a given drawing d can be based on the number of segments it contains. This is represented by $|S^m|$ in the above equation. This metric is intrinsic to the drawing itself, thereby avoiding dependencies on the length or verbosity of the instructions used to generate it, such as in a descriptive language like TikZ.

We also note here that while our implementation of teaching size is grounded in segment count for drawings, the framework itself is more general. Teaching size, as a proxy for descriptive complexity, can be adapted to other domains using modality-appropriate metrics.

3.2 Concepts

In our work, if the expected concept is `car` and the identified concept is `police car`, the identification is still considered correct because `police car` is a specific type of `car`, i.e., it is a semantically related prediction. This approach is similar to the one followed by Lamb et al. (2020). This means that if a specific sub-concept, or *hyponym*, is identified, it should still be seen as a correct identification as long as it falls under the more general expected concept. For a concept c, such as `car`, we consider a set of hyponyms $h(c)$ that corresponds to a set of concepts with a more specific meaning than c, e.g., `police car` belongs to $h(\texttt{car})$. For this study, we want a set of concepts that ensures that in the set of their hyponyms, there is no overlap, i.e., for any two concepts c_i, c_j, we have $h(c_i) \cap h(c_j) = \emptyset$. This rules out certain pairs of concepts available in the *Quick, Draw!*, like `van` and `car`, and it enhances the clarity and robustness of the study. We thus select the following subset of 20 concepts from the 345 concepts available in *Quick, Draw!*, with no overlap among their hyponyms: `apple`, `banana`, `car`, `cat`, `computer`, `cup`, `door`, `envelope`, `fish`, `grass`, `hockey puck`, `house`, `key`, `radio`, `string bean`, `sun`, `sword`, `television`, `The Great Wall of China` and `tree`.

In Table 3 in the Appendix [9], we list each concept from the dataset and the accepted hyponyms that are considered correct. This correspondence is established by human inspection and after the execution of the drawing selection phase (cf. Sect. 3.6) and the machine teaching framework experiments, with the results then analyzed based on these mappings.

3.3 Drawings

After choosing the concepts to study, we only include drawings that the game's neural network correctly identified (i.e., $r = 1$) in our research. For every concept, approximately 50 drawings are selected by a proportional random stratified sampling method [27], which groups drawings into bins based on their number of segments. (This number is approximate, as there may be rounding errors when calculating the number of samples for each bin according to its proportion.) The bin width was obtained using the minimum bin width between the Sturges's rule and the Freedman–Diaconis Estimator, ensuring that drawings of any concept are represented in a way that reflects the distribution of stroke counts for all correctly identified drawings of that concept in the dataset.

To simplify the drawings in our study, we employ the Ramer–Douglas–Peucker (RDP) algorithm [22, 7] on each stroke s of a given drawing d. RDP reduces the number of segments in each stroke while preserving its overall shape. Specifically, given a stroke s with a sequence of points $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, the RDP algorithm iteratively selects the most distant point (x_d, y_d) from the line segment connecting the first and last points of the stroke. If this distance is below a predefined threshold ϵ, then this stroke is simplified to a single segment $\{(x_1, y_1), (x_n, y_n)\}$ on the first and last points. However, if the distance to (x_d, y_d) exceeds ϵ, the algorithm keeps this point and recursively processes the two sequences of points formed by $\{(x_1, y_1), \ldots, (x_d, y_d)\}$ and $\{(x_d, y_d), \ldots, (x_n, y_n)\}$. This ensures that the essential characteristics of the stroke, up to distance ϵ, are preserved. This process continues until all points in the stroke fall within the threshold, resulting in a simplified representation of the stroke with fewer segments. By incrementing the threshold parameter, from an initial value of $\epsilon = 2$ [2], until each stroke is reduced to one segment, we generate simplified versions of each original drawing associated with a given concept c, resulting in new drawings $\{d\}_\epsilon \subseteq D_c$. Figure 2 illustrates a drawing simplification.

We note here that image and coordinate representations are generated differently, but both encode the same visual information. While not equivalent in all respects, the coordinates in TikZ are a form of structured data that, by reflecting a sequence of drawing actions, yield the same shape as the image once rendered.

Figure 2: Example of a drawing simplification for the concept `car` using the RDP algorithm. As the value of ϵ increases, the drawings become progressively simpler.

[2] The strokes stored in the Simplified Drawing files of *Quick, Draw!* have already been simplified by the RDP algorithm using $\epsilon = 2$, so this initial value did not simplify any drawing further.

3.4 Learners (L)

We utilize multiple LLMs, including two GPT-4 models (`gpt-4-turbo` and `gpt-4o`) from OpenAI, Llama (`Llama-3.2-90b-vision-instruct`) from Meta, Gemini (`gemini-pro-1.5`) from Google DeepMind, Pixtral (`pixtral-large-latest`) from Mistral, and Claude (`claude-3-5-sonnet`) from Anthropic. These models are capable of processing visual and language inputs to produce text outputs. To conduct the experiments of this work, all models were accessed via their respective Application Programming Interfaces (APIs). Additionally, we set the temperature parameter T to 1 for the experiments carried out within the machine teaching framework, and we set $T = 0$ for the drawing selection phase. $T \in [0..2]$ controls the behavior of the models' outputs: the lower T is, the more deterministic (predictable) results it leads to [20]. Thus, by setting $T = 0$ in the drawing selection phase, our goal is to obtain deterministic and predictable results, which are essential for creating a consistent baseline of drawings where the concepts were correctly identified. On the other hand, setting $T = 1$ in the experiments of the machine teaching framework is intended to introduce a controlled level of variability.

We consider two different representations for each concept: a visual representation and a text-based representation. Accordingly, we develop and test two prompt templates, one for each modality. For the vision-based modality, the drawings are presented as images generated from the sequence of coordinates (cf. Prompt 1 in the Appendix [9]). For the text-based modality, the pen stroke vectors are coded using the TikZ language (cf. Prompt 2 in the Appendix [9]). Both prompts ask for an open-ended answer (not multiple choice), allowing the learners to consider a wide range of possible concepts when identifying a given concept, including any that is not in our 20-concept set.

Data contamination occurs when language models are tested and evaluated using information from their training data, such as drawings already seen during training [23]. However, in this study, the drawings are consistently simplified using the RDP algorithm. This algorithm alters the coordinate information, thereby modifying the TikZ code and the visual representation. Consequently, we argue that these modified drawings are not part of the training set used to train the learners. Therefore, contamination tests are not required for this experiment.

It is important to note that although the models are not trained during our experiments, we refer to them as "learners", since this is aligned with the standardized terminology of machine teaching.

3.5 Concept Priors

As we argue in the introduction, some concepts, such as a `house`, are more common than others, such as an `envelope`. This sets a strong prior bias, especially in cases of doubt. For each of the 20 concepts, we use the 2022 English corpus of Google Books Ngram [12], providing the prior of a given concept as a normalized number between 0 and 1, representing the relative frequency of the concept. The rationale for using word frequency from Google Books Ngram as a proxy for human priors lies in the historical and cultural representativeness of a corpus. The assumption underlying our approach is that the frequency of specific words and phrases in written text correlates with their prominence in human thoughts, discussions, and collective knowledge at particular times [28]. Given that LLMs are trained on large text corpora that include books, articles, and other written materials, it is reasonable to assume that the Google Books Ngram priors closely align with the priors embedded in LLMs.

The priors were obtained in a case-insensitive manner. Each concept is treated exclusively as a noun to prevent confusion with its verb form (i.e., `fish` is interpreted as the animal and not the fishing activity).

3.6 Drawing Selection Phase

Before applying the machine teaching framework, we first conduct a drawing selection phase. This process identifies which drawings are reliably recognized by each model across modalities. These filtered examples form the basis for estimating teaching size. Hence, our minimization of Eq. 1 is sufficiently accurate.

As already mentioned, the drawings are simplified using the RDP algorithm, starting with a threshold of $\epsilon = 2$ on the raw drawings and continuing until each stroke in the drawing consists of a single segment. For each ϵ, the learner is prompted using Prompt 1 for visual-based identification and Prompt 2 for text-based identification (cf. Appendix [9]). Then, based on the completions from the learner, we obtain, by human inspection, the correspondence (between concepts and their respective accepted hyponyms) described in Table 3 in the Appendix [9], and we analyze the results based on those mappings. The accuracy and frequency of mistakes for each concept are obtained from the drawing selection phase.

In total, for the drawing selection phase, we run tests on each learner separately, generating a total of $21,896$ prompts—half $(10,948)$ for coordinates and half for images. These prompts were checked by human visual inspection, producing Table 3 (Appendix [9]). We then use the drawings that are correctly identified to test and evaluate the machine teaching framework proposed in Eq. 1, and thus obtain, for each concept, the teaching size.

4 Results

4.1 Concepts Identified

Out of the 20 concepts evaluated, all were identified in the image-based modality by at least one model. However, for the coordinates representation, `television`, `sword`, `radio`, `car`, `door`, `hockey puck`, `string bean`, and `The Great Wall of China` were never recognized by any model. We hypothesize that not only the complexity but also the prior of each of these latter concepts is behind their failed identification.

The image-based modality is thus more effective than the coordinate-based modality in identifying a broader range of concepts. This observation aligns with the typical human learning patterns, where visual information is often easier to process and understand than abstract textual-numerical data.

4.2 Accuracy

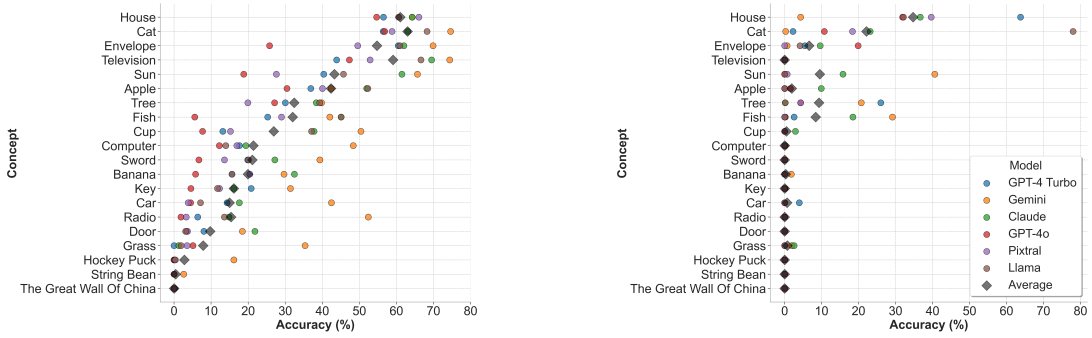


Figure 3: Accuracy for each concept in the vision-based (images; left) and text-based (coordinates; right) modality representations. ♦ represents the average accuracy value for the concept.

We begin by evaluating the accuracy on each concept c, Accuracy(c), defined here as

$$\text{Accuracy}(c) = \frac{1}{N_c} \sum_{i=1}^{N_c} \mathbb{1}\left[L(S_i) = c\right], \tag{2}$$

where N_c corresponds to the total number of tests (in this case, prompts) conducted on L for the concept c on the drawing selection phase, with $\{S_i\}_{i=1}^{N_c} \subseteq D_c$.

Figure 3 depicts each concept's accuracy across the two modality representations. The image modality shows a wider range of recognition accuracy, with average performance metrics spanning up to 65%. In contrast, the coordinate modality exhibits a much narrower range, largely confined to 0–25% average accuracy. This discrepancy likely reflects the models' ability to leverage richer visual features in image-based representations compared to the sparse and abstract nature of coordinate-based inputs. The richer detail in images provides more cues for concept identification, while the textual coordinates impose a more constrained and abstract recognition task.

Nevertheless, the results suggest that some concepts are fundamentally challenging to recognize, regardless of the modality. `House` and `cat` achieve relatively high accuracy across both modalities, indicating their simplicity or recognizability regardless of representation. In contrast, more complex or less visually distinct concepts, such as `hockey puck` and `The Great Wall of China`, show zero accuracy in the coordinate modality and only marginal performance in the image modality.

Among the models evaluated, Gemini emerges as the best-performing model in the image modality, consistently achieving better results across a broader range of concepts. Notably, it stands apart from the other models, which appear to form a distinct cluster in terms of performance. This suggests that while certain concepts are uniformly challenging

across all models, Gemini is better equipped to handle a wider range of visual representations. In the coordinate modality, no single model shows clear superiority, likely due to the shared constraints of the textual representation.

The precise accuracy of each concept, categorized by model and modality, can be found in Table 4 in the Appendix [9].

We also study the relationship between the number of segments (i.e., complexity) and the accuracy of concept identification for both image- and coordinate-based representations, as shown in Figure 4. For image-based representations, there is a clear positive relationship between accuracy and the number of segments. Starting from an accuracy of around $0.3\,\%$ in the $(0, 4]$ interval, the accuracy increases steadily, reaching approximately $50\,\%$ in the $(29, 69]$ interval.

Conversely, for coordinate-based representations, the average accuracy remains significantly lower and follows a more modest increasing trend. Beginning at roughly $1\,\%$, it gradually rises to around $8\,\%$ in the $(16, 19]$ interval before stabilizing and fluctuating slightly in the higher segment intervals. This indicates that increasing the number of segments in coordinate-based representations provides only minimal benefits in accuracy.

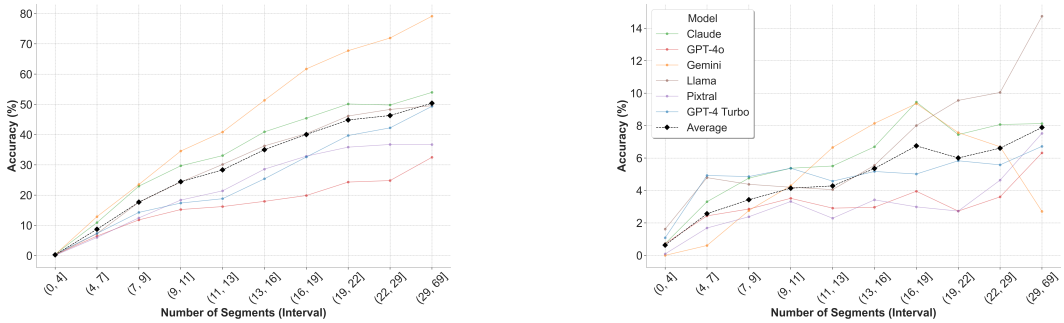


Figure 4: Relationship between the number of segments and accuracy for both modalities (images; left) (coordinates; right).

4.3 Frequency of Mistakes

Accuracy measures how well the learner has identified the correct concepts. However, the model can also respond with "I don't know" answers (or something that is not a concept) or by identifying a different concept that is incorrect. We focus on the latter case and refer to this performance metric as the *frequency of mistakes* for a given concept c in model m, $\text{FOM}_m(c)$.

Formally,

$$\text{FOM}_m(c) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\left[S_i \notin D_c \wedge L_m(S_i) = c\right], \tag{3}$$

where $N = 10,948$ is the total number of tests (prompts) conducted on L during the drawing selection phase on each modality.

To simplify the interpretation of the frequency of mistakes for a given concept, we average the $\text{FOM}(c)$ across all models. We also explore whether there is a relationship between the frequency of mistakes and the prior probability of each concept. We have included in Tables 8 to 19 of the Appendix [9] the confusion matrices for each model and modality. These tables show how well the model performs across various concepts by detailing the true positives and the frequency of errors for each concept. Figure 5 shows that the vision modality exhibits a lower percentage of observed mistakes than the coordinate-based modality.

Interestingly, as shown in Figure 3, the concept `house` in both modality representations, `television` only in the visual-based modality, and `cat` only in the text-based modality, shows the highest accuracy. However, these concepts also have the highest frequency of mistakes, indicating that while they are often correctly identified, they are also frequently guessed when wrong. This indicates that although these concepts are generally easily recognizable, variations in attributes like size and shape may introduce ambiguities that complicate the identification of these concepts. In other words, the models often guess these concepts, whether they are correct or not.

When calculating the Pearson correlation between the frequency of mistakes and the prior probability, we obtain a correlation of 0.914 for the coordinate-based modality and 0.434 for the vision-based modality for all concepts. This suggests that in the textual modality, the learner is more susceptible to responding based on their pre-existing biases when confronted with unfamiliar concepts. In contrast, this tendency is reduced in visual representation.

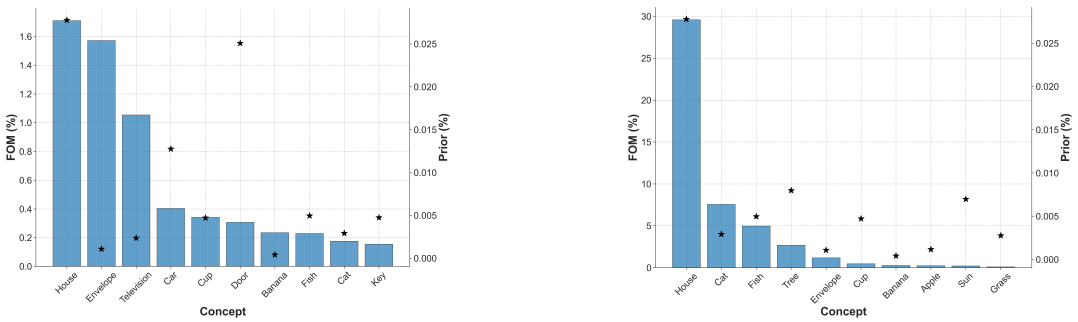


Figure 5: Top-10 concepts with the highest frequency of mistakes (averaged across the models) in the visual-based modality (images) (left) and text-based modality (coordinates) (right). The little star represents the prior probability for each concept.

4.4 Teaching Size

To calculate the teaching size for each concept, we set T to 1, ρ to 0.5, and N to 50, meaning that a correct identification needs to happen at least 25 times out of 50 trials even with some stochasticity in the model. The aim is to determine the simplest drawing for each modality representation that the learner can identify consistently in at least 25 out of 50 trials. We highlight that this procedure is different from the one conducted in the previous sections, where the results came from the drawing selection phase.

We present the results for teaching size of images and coordinates in Tables 5 and 6 in the Appendix [9]. Table 7 of the Appendix [9] shows the respective simplest drawings identified for each concept, modality and model. The data suggest that, on average, the teaching size values for coordinates (11.46, SD=8.60) with successful identification (12) are higher than those for images (6.73, SD=2.25) with successful identification (20), regardless of the model. Even when considering only the 12 concepts that are well identified using coordinates, the mean teaching size remains lower for images. This indicates that there is no absolute invariance, answering our question Q1 in the negative. In other words, the number of strokes required for a concept to be identified by the learners is generally higher when using textual coordinates compared to bitmap images.

Table 2: Concept teaching size comparison for images and for coordinates, showing Kendall Rank correlation coefficient for the subset of concepts that are identified (*), and Pearson correlation between the accuracy for all concepts.

Model	Order for Images	Order for Coordinates	Rank*	Pears
Claude	`cup` < `house` = `fish` < `envelope` < `apple` = `sun` < `cat` < `grass`	`house` < `envelope` < `apple` < `fish` < `cat` < `sun` < `cup` < `grass`	0.36	0.65
Gemini	`envelope` = `house` = `sun` = `grass` = `fish` < `banana` < `tree` = `cat`	`envelope` < `house` < `sun` = `tree` < `fish` < `banana` < `grass` < `cat`	0.57	0.21
GPT-4o	`tree` < `house` < `envelope` < `apple` < `cat`	`envelope` < `house` < `cat` < `tree` < `apple`	0.00	0.67
GPT-4T	`envelope` < `house` < `fish` < `cat` < `tree` < `car`	`envelope` = `house` < `fish` = `tree` < `cat` < `car`	0.87	0.45
Llama	`envelope` = `house` < `cat`	`envelope` = `house` < `cat`	1.00	0.50
Pixtral	`house` < `cat`	`house` < `cat`	1.00	0.63

Furthermore, it is important to highlight a weak, though similar, negative correlation between the teaching size and the prior of each concept across both modalities. The correlation coefficients are -0.021 for coordinates and -0.338 for images, over all concepts and models. This suggests that, in the image modality, the more common a concept is, the simpler its drawings need to be for the learner to consistently identify it.

Interestingly, looking at Table 2, the teaching size still ranks concepts in a relatively similar order between images and coordinates, but the strength of this relationship varies across models. The strongest agreement is observed in Llama and Pixtral, both of which exhibit a perfect Kendall rank correlation of 1.0, meaning their rankings are identical across the two modalities. GPT-4 Turbo (shortened as GPT-4T) also exhibits a high correlation (0.87), suggesting strong alignment in concept difficulty ordering between images and coordinates.

However, other models show lower correlations, with Claude at 0.36, Gemini at 0.57, and GPT-4o displaying no correlation (0.0) between the two rankings. To control for the influence of concept priors on teaching size, we performed an ordinary least squares regression of the teaching sizes (for both modalities) on the corresponding concept priors derived from Google Books Ngram frequencies. This yielded residuals representing the portion of teaching size not explained by prior familiarity. We then calculated the Kendall rank correlation between these residuals from different modalities and found similar correlation values.

These results indicate that while some models maintain an invariant notion of teaching size across modalities, others exhibit some discrepancies, although the number of concepts is small. The accuracy correlation between all concepts is a more robust metric, and it also calculates how well concept-wise accuracies align between the two modalities. Claude and GPT-4o exhibit relatively high accuracy correlations (0.65 and 0.67, respectively), suggesting that despite their lower Kendall rank correlations, the overall accuracy patterns remain similar. Meanwhile, Gemini and GPT-4T have lower accuracy correlations (0.21 and 0.45), compensated by the better values for ranking.

Overall, the correlations are never negative, but less or more positive depending on the model. While Llama, Pixtral, and GPT-4T exhibit strong invariance in teaching size ranking across modalities, others do not. In general, however, the answer to question Q2 tends to be positive.

5 Discussion

In this study, we examined how multimodal models identify the same concepts in two different modalities: image- and coordinate-based drawings. Our findings show that images are generally more effective than coordinates for identifying concepts. In particular, using images led to the recognition of more concepts than when using coordinates, indicating that images are better suited for teaching concepts to a given learner. This is supported by the higher accuracy and lower frequency of mistakes seen with image-based representations. We also use the number of segments as the teaching size to measure the complexity of a concept. Our analysis indicates that the teaching size is again more beneficial for images than coordinates (clearly answering question Q1 negatively), but ranks concepts in similar ways, regardless of the type of drawing used, even when we account for the learner's priors. While there are differences depending on the model, we tend to see a positive answer to question Q2 more often. This suggests that some concepts are naturally easier or more difficult to teach, no matter how they are represented.

We believe that our study provides a step towards the investigation of a core question in the field of multimodal Artificial Intelligence (AI): Whether language models can interpret structured data (like coordinates) as effectively as images. We saw that models perform better with image-based representations, even for simple concepts. This suggests a limitation in current multimodal models that is important for scientific and practical development. The observed invariance in ranking teaching size across modalities suggests that some concept properties are robust regardless of representation. This may help improve cross-modal transfer learning, where models must generalize concepts between formats.

Our machine teaching framework has several practical implications. First, it improves the design and evaluation of multimodal systems by providing a quantitative, model-agnostic way to measure how "costly" it is for any vision-language model to learn a new visual concept in different modalities. Second, our work connects cognitive and computational notions of simplicity by providing empirical evidence that segment count, a classic cognitive cue, remains predictive even in state-of-the-art large language models. This contributes to ongoing discussions in AI about whether these models learn conceptual structures or simply memorize patterns. Finally, the framework can support adaptive teaching tools by identifying the simplest representations for individual learning needs. Thus, it could be used to develop educational software that teaches geometric concepts or visual reasoning using minimal and optimally chosen examples.

Our analysis has to be seen in the light of some limitations. (a) The study concentrates on a specific set of concepts, which might affect how well the findings apply to other (potentially more complex) concepts. (b) Our use of the RDP algorithm for drawing simplification streamlines each stroke but does not totally remove any single stroke from the drawing. This should not be much of a limitation as we focus on the simplest concepts. (c) A factor that can influence the teaching size of a concept is the curvature of its drawings, i.e., the amount by which it deviates from a straight line. In this work, we have chosen not to focus on this aspect, but this could be of interest for future works.

We show that the simplest concepts usually correspond to those that humans intuitively think of as less complex, and this confirms that the simplest concepts are so across modalities. This supports the hypothesis that the representation of concepts in both modalities is tightly connected in the latent space. However, since we operate under a black-box setting with models like GPT-4 and others that do not expose their internal representations, we cannot directly inspect or confirm such latent alignments. Some other methods, especially white-box approaches that have access to weights or gradients, could give a definitive answer to this hypothesis. Still, in cases such as GPT-4 or humans, a black-box

approach such as the one presented in this paper is the practical course of action. Thus, our results should be viewed as a hypothesis to explain the invariance across modalities and not as a definitive claim.

The code to reproduce our results is available [10].

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, et al. GPT-4 technical report, 2023.

[2] Bajaj, Payal. The Quick, Draw! - A.I. `https://github.com/payalbajaj/sketch_rnn_classification`, 2017. Accessed: 2024-08-02.

[3] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, et al. Sparks of artificial general intelligence: Early experiments with GPT-4, 2023.

[4] Mu Cai, Zeyi Huang, Yuheng Li, Utkarsh Ojha, Haohan Wang, et al. Leveraging large language models for scalable vector graphics-driven image understanding, 2023.

[5] May Jane Chen and Michael Cook. Representational drawings of solid objects by young children. *Perception*, 13 (4):377–385, 1984. doi: 10.1068/p130377.

[6] Thorsten Doliwa, Gaojian Fan, Hans Ulrich Simon, and Sandra Zilles. Recursive teaching dimension, VC-dimension and sample compression. *The Journal of Machine Learning Research*, 15(1):3107–3131, 2014. doi: 10.5555/2627435.2697064.

[7] David H Douglas and Thomas K Peucker. Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *Cartographica: The International Journal for Geographic Information and Geovisualization*, 10(2):112–122, 1973. doi: 10.3138/FM57-6770-U75U-7727.

[8] Mathias Eitz, James Hays, and Marc Alexa. How do humans sketch objects? In *Proceedings of the 2012 ACM Special Interest Group on Computer Graphics and Interactive Techniques (SIGGRAPH)*, volume 31, pages 1(44)–10(44), Los Angeles, CA, USA, 2012. doi: 10.1145/2185520.2185540.

[9] Diogo Freitas, Brigt Håvardstun, Cèsar Ferri, Darío Garigliotti, Jan Arne Telle, and Jose Hernandez-Orallo. Relative drawing identification complexity is invariant to modality in vision-language models, 2025. Full version of this paper.

[10] Diogo Nuno Freitas, Brigt Håvardstun, Dario Garigliotti, Jan Arne Telle, Cèsar Ferri Ramírez, and Jose Hernandez-Orallo. Code for the paper: "Relative Drawing Identification Complexity is Invariant to Modality in Vision-Language Models", 2025. Available at: `https://doi.org/10.5281/zenodo.16762246`.

[11] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, et al. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness, 2023.

[12] Google. Google Ngram Viewer. `https://books.google.com/ngrams/`, 2010. Accessed: 2024-07-09.

[13] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, et al. The Llama 3 herd of models, 2024.

[14] David Ha and Douglas Eck. A neural representation of sketch drawings, 2017.

[15] Jonas Jongejan, Henry Rowley, Takashi Kawashima, Jongmin Kim, and Nick Fox-Gieg. The Quick, Draw! - A.I. `https://quickdraw.withgoogle.com/`, 2016. Accessed: 2024-07-08.

[16] Abdullah Talha Kabakus. A novel sketch recognition model based on convolutional neural networks. In *Proceedings of the 2020 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*, pages 1–6, Ankara, Turkey, 2020. doi: 10.1109/HORA49412.2020.9152911.

[17] Alex Lamb, Sherjil Ozair, Vikas Verma, and David Ha. SketchTransfer: A new dataset for exploring detail-invariance and the abstractions learned by deep networks. In *Proceedings of the 2020 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 952–961, Snowmass Village, CO, USA, 2020. doi: 10.1109/WACV45572.2020.9093327.

[18] Bria Long, Judith E Fan, and Michael C Frank. Drawings as a window into developmental changes in object representations. In *Proceedings of the 40th Annual Conference of the Cognitive Science Society*, pages 708–713, Madison, WI, USA, 2018. doi: 10.1167/18.10.398.

[19] Shay Moran and Amir Yehudayoff. Sample compression schemes for VC classes. *Journal of the ACM*, 63(3): 1(21)–10(21), 2016. doi: 10.1145/2890490.

[20] OpenAI. API reference. `https://platform.openai.com/docs/api-reference`, 2024. Accessed: 2024-07-09.

[21] Reza Pourreza, Apratim Bhattacharyya, Sunny Panchal, Mingu Lee, Pulkit Madan, et al. Painter: Teaching auto-regressive language models to draw sketches. In *Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 305–314, Paris, France, 2023. doi: 10.1109/ICCVW60793.2023.00038.

[22] Urs Ramer. An iterative procedure for the polygonal approximation of plane curves. *Computer Graphics and Image Processing*, 1(3):244–256, 1972. doi: 10.1016/S0146-664X(72)80017-0.

[23] Mathieu Ravaut, Bosheng Ding, Fangkai Jiao, Hailin Chen, Xingxuan Li, et al. How much are LLMs contaminated? A comprehensive survey and the LLMSanitize library, 2024.

[24] Rosália G. Schneider and Tinne Tuytelaars. How do humans sketch objects? In *Proceedings of the 2014 ACM Special Interest Group on Computer Graphics and Interactive Techniques (SIGGRAPH) Asia*, volume 33, pages 1(174)–9(174), Shenzhen, China, 2014. doi: 10.1145/2661229.2661231.

[25] Pratyusha Sharma, Tamar Rott Shaham, Manel Baradad, Stephanie Fu, Adrian Rodriguez-Munoz, et al. A vision check-up for language models, 2024.

[26] Zekun Sun and Chaz Firestone. Seeing and speaking: How verbal "description length" encodes visual complexity. *Journal of Experimental Psychology: General*, 151(1):82–96, 2022. doi: 10.1037/xge0001076.

[27] Hamed Taherdoost. Sampling methods in research methodology: How to choose a sampling technique for research. *International Journal of Academic Research in Management*, 5(2):18–27, 2016. doi: 10.2139/ssrn.3205035.

[28] Kumiko Tanaka-Ishii and Hiroshi Terada. Word familiarity and frequency. *Studia Linguistica*, 65(1):96–116, 2011. doi: 10.1111/j.1467-9582.2010.01176.x.

[29] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024.

[30] Jan Arne Telle, José Hernández-Orallo, and Cèsar Ferri. The teaching size: Computable teachers and learners for universal languages. *Machine Learning*, 108(8–9):1653–1675, 2019. doi: 10.1007/s10994-019-05821-2.

[31] Yael Vinker, Tamar Rott Shaham, Kristine Zheng, Alex Zhao, Judith E Fan, et al. SketchAgent: Language-driven sequential sketch generation, 2024.

[32] Peng Xu, Chaitanya K. Joshi, and Xavier Bresson. Multigraph transformer for free-hand sketch recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 33(10):5150–5161, 2022. doi: 10.1109/TNNLS.2021.3069230.

[33] Fan Yang, Nor Azman Ismail, Yee Yong Pang, Victor R Kebande, Arafat Al-Dhaqm, et al. A systematic literature review of deep learning approaches for sketch-based image retrieval: datasets, metrics, and future directions. *IEEE Access*, 12:14847–14869, 2024. doi: 10.1109/ACCESS.2024.3357939.

[34] Qian Yu, Yongxin Yang, Yi-Zhe Song, Tao Xiang, and Timothy Hospedales. Sketch-a-Net that beats humans, 2015.

[35] Xingyuan Zhang, Yaping Huang, Qi Zou, Yanting Pei, Runsheng Zhang, et al. A hybrid convolutional neural network for sketch recognition. *Pattern Recognition Letters*, 130:73–82, 2020. doi: 10.1016/j.patrec.2019.01.006.

[36] Xiaojin Zhu, Adish Singla, Sandra Zilles, and Anna N. Rafferty. An overview of machine teaching, 2018.

A Prompts Utilized in this Study

In this work, we use two prompt templates to evaluate the effectiveness of concept identification across multiple models–GPT-4, Llama, Gemini, Pixtral, and Claude–using two different modalities: vision-based and text-based representations.

For the visual modality, the drawings are presented as images generated from the sequence of coordinates. The prompt template for this modality involves showing the model the bitmap image of the drawing and asking it to identify the concept depicted in the image.

For the textual modality, the pen stroke vectors are encoded using the TikZ language. This format allows the representation of drawings as a series of coordinates and commands that describe the strokes. The prompt template for this modality involves presenting the model with these TikZ-encoded coordinates and asking it to identify the concept represented by the strokes.

Both prompts are designed to elicit open-ended responses from the model, allowing it to consider a wide range of possible concepts, including those not in the predefined 20-concept set. This approach ensures that the model's identification process is not constrained by a limited set of options, thereby providing a more comprehensive evaluation of its capabilities in both modalities.

Prompt 1: Prompt template for the vision-based modality.

```
 Your task is to identify a concept drawn by hand.  You will be provided with an image
corresponding to a concept drawn by hand.  Your task is to identify, based on the provided
picture, the concept that someone has attempted to draw.  Please reply only with the name of
the concept.
```

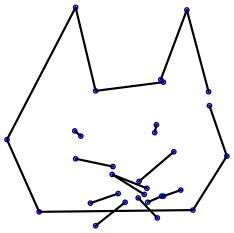
Image URL: base 64 encoded drawing (256×256)

Prompt 2: Prompt template for the text-based modality.

```
 Your task is to identify a concept drawn by hand.  You will be provided a TikZpicture
format corresponding to a concept, where each stroke is indicated by the command 'draw'
followed by a series of points in '(x,y)' format.
The points are connected by straight lines, denoted by '--'.  The strokes collectively
represent a concept.  Below is the TikZpicture code enclosed within triple backticks:
'''{TikZ code}'''.
Your task is to identify, based on the provided TikZpicture, the concept that someone has
attempted to draw.  Please reply only with the name of the concept.
```

Example of image for the concept `cat` The vision-based modality, on the other hand, involves using images created from the sequence of coordinates from the *Quick, Draw!*dataset. These images are produced by plotting the coordinates with a function that defines the image size as 256×256 pixels. The image is then stored in PNG format.

The following is an example of an image representing the `cat`, extracted from the *Quick, Draw!* dataset.



Example of TikZ code for the concept `cat` TikZ is a LaTeX package used for creating graphics programmatically. Because of its way of representing drawings through coordinate-based commands, we used TikZ in the text-modality tests.

Each drawing in the *Quick, Draw!* dataset is stored as vectors of distinct pen strokes, represented by sequences of (x, y) coordinates. For each stroke in the drawing, the sequence of points is translated into a `\draw` command. The points are connected using the `--` operator, which denotes a straight line between two points. Each drawing consists of multiple strokes, and each segment is represented by a separate `\draw` command in TikZ.

The following is an example of the TikZ code of the concept `cat`, extracted from the *Quick, Draw!* dataset.

```
\draw (181, 30) -- (121, 12) -- (14, 95) -- (0, 161) -- (42, 255) --
    (73, 213) -- (136, 226) -- (236, 194) -- (242, 230) -- (255, 156) --
    (218, 38) -- (161, 2) -- (141, 15);
\draw (118, 92) -- (76, 118);
\draw (119, 81) -- (87, 76);
\draw (112, 70) -- (102, 57);
\draw (146, 98) -- (192, 107);
\draw (151, 76) -- (203, 86);
\draw (154, 53) -- (175, 51);
\draw (135, 138) -- (137, 71) -- (123, 81);
```

B Accepted Hyponyms for each Concept

Table 3: Accepted hyponyms for each concept. In this study, we establish a set of accepted hyponyms for each concept. A hyponym is a more specific term within a broader category, and for our purposes, identifying a hyponym is considered correct if it falls under the general expected concept. For instance, if the expected concept is `car`, identifying `ambulance` is still correct because it is a specific type of car. This table lists each concept and its accepted hyponyms. These hyponyms are identified in the drawing selection phase and validated by human inspection.

Concept (c)	Hyponyms ($h(c)$)
Apple	Apple logo
Banana	Banana peel Banana pepper Banana/crescent moon
Car	Ambulance Truck Pickup truck Tractor Tank
Cat	Cat whiskers Cat face Cat head Cat playing with a ball of yarn Cat playing with a toy Cat/fox House with a cat A cat chasing a mouse
Computer	Laptop Desktop computer
Cup	Glass Broken cup Broken glass Coffee cup Coffee mug Cup and saucer Cup of coffee Cup/glass Glass and napkin Glass of water Jar Jug Mug Pitcher Wine glass
Door	Car door Door with a doorknob Doorway Door with a handle Door ajar Swinging door
Envelope	(no hyponyms from the completions)
Fish	Whale
Grass	Grass/sawtooth wave
Hockey puck	Hockey puck and stick
House	Triangle and house
Key	Key and knife

Continued on next page.

Concept (c)	Hyponyms ($h(c)$)
Radio	Radio controller Radio controlled car Radio cassette player
String bean	(no hyponyms from the completions)
Sun	Sunburst Starburst Sun rays Sun/star
Sword	Sword in the stone Knife Khukuri
Television	TV Television/TV/monitor/screen Television/TV/monitor Line graph on a TV screen Computer monitor Monitor Desktop monitor Computer monitor
Tree	Palm tree Christmas tree Tree branch
The Great Wall of China	(no hyponyms from the completions)

C Accuracy for each Concept, Model and Modality

Table 4: We evaluate the accuracy of concept identification in the drawing selection phase. Accuracy is determined by the proportion of correctly identified concepts over the total number of evaluations for each concept. A correct identification includes cases where a hyponym of the expected concept is recognized, as established in our accepted hyponym mappings. This table presents the accuracy scores for each concept and model, based on responses from the learners when presented with drawings in both visual and text-based representations.

Concept (c)	Model (m)	Modality	Accuracy (%)
The Great Wall Of China	Claude	Images	0.00
	Claude	Coordinates	0.00
	GPT-4 Turbo	Images	0.00
	GPT-4 Turbo	Coordinates	0.00
	GPT-4o	Images	0.18
	GPT-4o	Coordinates	0.00
	Gemini	Images	0.00
	Gemini	Coordinates	0.00
	Llama	Images	0.00
	Llama	Coordinates	0.00
	Pixtral	Images	0.00
	Pixtral	Coordinates	0.00
String Bean	Claude	Images	0.00
	Claude	Coordinates	0.00
	GPT-4 Turbo	Images	0.00
	GPT-4 Turbo	Coordinates	0.00
	GPT-4o	Images	0.00
	GPT-4o	Coordinates	0.00
	Gemini	Images	2.61
	Gemini	Coordinates	0.00
	Llama	Images	0.00
	Llama	Coordinates	0.00
	Pixtral	Images	0.00
	Pixtral	Coordinates	0.00
Hockey Puck	Claude	Images	0.00
	Claude	Coordinates	0.00
	GPT-4 Turbo	Images	0.00
	GPT-4 Turbo	Coordinates	0.00
	GPT-4o	Images	0.00
	GPT-4o	Coordinates	0.00
	Gemini	Images	16.14
	Gemini	Coordinates	0.00
	Llama	Images	0.37
	Llama	Coordinates	0.00
	Pixtral	Images	0.19
	Pixtral	Coordinates	0.00
Grass	Claude	Images	1.27
	Claude	Coordinates	2.55
	GPT-4 Turbo	Images	0.00
	GPT-4 Turbo	Coordinates	0.00
	GPT-4o	Images	5.10
	GPT-4o	Coordinates	0.00
	Gemini	Images	35.35
	Gemini	Coordinates	1.91
	Llama	Images	1.91
	Llama	Coordinates	0.00
	Pixtral	Images	3.50

Continued on next page.

Concept (c)	Model (m)	Modality	Accuracy (%)
	Pixtral	Coordinates	0.00
Door	Claude	Images	21.82
	Claude	Coordinates	0.00
	GPT-4 Turbo	Images	8.05
	GPT-4 Turbo	Coordinates	0.00
	GPT-4o	Images	3.38
	GPT-4o	Coordinates	0.00
	Gemini	Images	18.44
	Gemini	Coordinates	0.00
	Llama	Images	3.12
	Llama	Coordinates	0.00
	Pixtral	Images	3.64
	Pixtral	Coordinates	0.00
Radio	Claude	Images	14.82
	Claude	Coordinates	0.00
	GPT-4 Turbo	Images	6.40
	GPT-4 Turbo	Coordinates	0.00
	GPT-4o	Images	1.87
	GPT-4o	Coordinates	0.00
	Gemini	Images	52.42
	Gemini	Coordinates	0.00
	Llama	Images	13.57
	Llama	Coordinates	0.00
	Pixtral	Images	3.28
	Pixtral	Coordinates	0.00
Car	Claude	Images	17.60
	Claude	Coordinates	0.00
	GPT-4 Turbo	Images	14.29
	GPT-4 Turbo	Coordinates	3.95
	GPT-4o	Images	4.46
	GPT-4o	Coordinates	0.00
	Gemini	Images	42.47
	Gemini	Coordinates	0.00
	Llama	Images	7.14
	Llama	Coordinates	0.00
	Pixtral	Images	3.83
	Pixtral	Coordinates	0.00
Key	Claude	Images	16.12
	Claude	Coordinates	0.00
	GPT-4 Turbo	Images	20.80
	GPT-4 Turbo	Coordinates	0.00
	GPT-4o	Images	4.55
	GPT-4o	Coordinates	0.00
	Gemini	Images	31.40
	Gemini	Coordinates	0.14
	Llama	Images	11.71
	Llama	Coordinates	0.00
	Pixtral	Images	12.26
	Pixtral	Coordinates	0.00

Continued on next page.

Concept (c)	Model (m)	Modality	Accuracy (%)
Banana	Claude	Images	32.46
	Claude	Coordinates	0.00
	GPT-4 Turbo	Images	15.63
	GPT-4 Turbo	Coordinates	0.00
	GPT-4o	Images	5.81
	GPT-4o	Coordinates	0.00
	Gemini	Images	29.66
	Gemini	Coordinates	1.80
	Llama	Images	15.63
	Llama	Coordinates	0.00
	Pixtral	Images	20.44
	Pixtral	Coordinates	0.00
Sword	Claude	Images	27.20
	Claude	Coordinates	0.00
	GPT-4 Turbo	Images	19.87
	GPT-4 Turbo	Coordinates	0.00
	GPT-4o	Images	6.69
	GPT-4o	Coordinates	0.00
	Gemini	Images	39.33
	Gemini	Coordinates	0.00
	Llama	Images	20.08
	Llama	Coordinates	0.00
	Pixtral	Images	13.60
	Pixtral	Coordinates	0.00
Computer	Claude	Images	19.37
	Claude	Coordinates	0.17
	GPT-4 Turbo	Images	17.63
	GPT-4 Turbo	Coordinates	0.00
	GPT-4o	Images	12.22
	GPT-4o	Coordinates	0.00
	Gemini	Images	48.34
	Gemini	Coordinates	0.00
	Llama	Images	13.96
	Llama	Coordinates	0.00
	Pixtral	Images	16.93
	Pixtral	Coordinates	0.00
Cup	Claude	Images	37.78
	Claude	Coordinates	2.91
	GPT-4 Turbo	Images	13.16
	GPT-4 Turbo	Coordinates	0.00
	GPT-4o	Images	7.69
	GPT-4o	Coordinates	0.00
	Gemini	Images	50.43
	Gemini	Coordinates	0.00
	Llama	Images	37.09
	Llama	Coordinates	0.00
	Pixtral	Images	15.21
	Pixtral	Coordinates	0.00
Fish	Claude	Images	45.08
	Claude	Coordinates	18.47
	GPT-4 Turbo	Images	25.25
	GPT-4 Turbo	Coordinates	2.54
	GPT-4o	Images	5.59
	GPT-4o	Coordinates	0.17
	Gemini	Images	42.03

Continued on next page.

Concept (c)	Model (m)	Modality	Accuracy (%)
	Gemini	Coordinates	29.15
	Llama	Images	45.08
	Llama	Coordinates	0.00
	Pixtral	Images	28.98
	Pixtral	Coordinates	0.00
Tree	Claude	Images	38.36
	Claude	Coordinates	0.16
	GPT-4 Turbo	Images	30.02
	GPT-4 Turbo	Coordinates	26.00
	GPT-4o	Images	27.13
	GPT-4o	Coordinates	4.33
	Gemini	Images	39.81
	Gemini	Coordinates	20.71
	Llama	Images	39.33
	Llama	Coordinates	0.16
	Pixtral	Images	19.90
	Pixtral	Coordinates	4.33
Apple	Claude	Images	51.96
	Claude	Coordinates	9.89
	GPT-4 Turbo	Images	36.89
	GPT-4 Turbo	Coordinates	0.00
	GPT-4o	Images	30.46
	GPT-4o	Coordinates	1.57
	Gemini	Images	42.39
	Gemini	Coordinates	0.00
	Llama	Images	52.28
	Llama	Coordinates	0.00
	Pixtral	Images	40.03
	Pixtral	Coordinates	0.00
Sun	Claude	Images	61.48
	Claude	Coordinates	15.78
	GPT-4 Turbo	Images	40.37
	GPT-4 Turbo	Coordinates	0.00
	GPT-4o	Images	18.79
	GPT-4o	Coordinates	0.00
	Gemini	Images	65.66
	Gemini	Coordinates	40.60
	Llama	Images	45.71
	Llama	Coordinates	0.00
	Pixtral	Images	27.61
	Pixtral	Coordinates	0.70
Television	Claude	Images	69.49
	Claude	Coordinates	0.00
	GPT-4 Turbo	Images	43.86
	GPT-4 Turbo	Coordinates	0.00
	GPT-4o	Images	47.29
	GPT-4o	Coordinates	0.00
	Gemini	Images	74.37
	Gemini	Coordinates	0.00
	Llama	Images	66.61
	Llama	Coordinates	0.00
	Pixtral	Images	52.89
	Pixtral	Coordinates	0.00

Continued on next page.

Concept (c)	Model (m)	Modality	Accuracy (%)
Envelope	Claude	Images	62.01
	Claude	Coordinates	9.61
	GPT-4 Turbo	Images	60.48
	GPT-4 Turbo	Coordinates	5.46
	GPT-4o	Images	25.76
	GPT-4o	Coordinates	19.87
	Gemini	Images	69.87
	Gemini	Coordinates	0.66
	Llama	Images	60.92
	Llama	Coordinates	4.15
	Pixtral	Images	49.56
	Pixtral	Coordinates	0.00
Cat	Claude	Images	63.05
	Claude	Coordinates	23.13
	GPT-4 Turbo	Images	56.42
	GPT-4 Turbo	Coordinates	2.26
	GPT-4o	Images	56.84
	GPT-4o	Coordinates	10.72
	Gemini	Images	74.61
	Gemini	Coordinates	0.28
	Llama	Images	68.27
	Llama	Coordinates	78.00
	Pixtral	Images	58.82
	Pixtral	Coordinates	18.34
House	Claude	Images	64.24
	Claude	Coordinates	36.67
	GPT-4 Turbo	Images	56.49
	GPT-4 Turbo	Coordinates	63.78
	GPT-4o	Images	54.67
	GPT-4o	Coordinates	31.89
	Gemini	Images	64.24
	Gemini	Coordinates	4.33
	Llama	Images	60.59
	Llama	Coordinates	32.12
	Pixtral	Images	66.06
	Pixtral	Coordinates	39.64

D Teaching size for each Concept, Model and Modality

Table 5: This table presents the teaching size (TS) for each concept in the image modality. The teaching size represents the minimal number of segments required in a drawing for a learner (i.e., a large language model) to recognize the concept with a probability of at least ρ over N independent trials. A lower teaching size indicates a simpler representation of the concept that is still consistently identifiable by the model.

Concept	Model	$\mathrm{TS}_{0.5,50}(c)$	Correct
Apple	Claude	7	50
	Gemini	9	50
	GPT-4 Turbo	9	50
	GPT-4o	7	43
	Llama	8	40
	Pixtral	10	31
Banana	Claude	8	50
	Gemini	8	50
	GPT-4 Turbo	10	50
	GPT-4o	15	45
	Llama	10	32
	Pixtral	9	32
Car	Claude	14	50
	Gemini	10	50
	GPT-4 Turbo	19	50
	GPT-4o	23	50
	Llama	18	27
	Pixtral	25	27
Cat	Claude	9	38
	Gemini	9	50
	GPT-4 Turbo	11	50
	GPT-4o	9	48
	Llama	9	33
	Pixtral	9	39
Computer	Claude	11	50
	Gemini	3	42
	GPT-4 Turbo	6	50
	GPT-4o	6	48
	Llama	10	34
	Pixtral	8	44
Cup	Claude	4	46
	Gemini	4	27
	GPT-4 Turbo	13	50
	GPT-4o	5	50
	Llama	6	34
	Pixtral	11	32
Door	Claude	4	37
	Gemini	5	50
	GPT-4 Turbo	7	45
	GPT-4o	4	47
	Llama	16	26
	Pixtral	10	36
Envelope	Claude	6	50
	Gemini	5	34
	GPT-4 Turbo	5	50
	GPT-4o	6	50
	Llama	6	42
	Pixtral	6	37

Continued on next page.

Concept	Model	$TS_{0.5,50}(c)$	Correct
Fish	Claude	5	44
	Gemini	5	50
	GPT-4 Turbo	9	50
	GPT-4o	9	46
	Llama	7	40
	Pixtral	6	42
Grass	Claude	14	33
	Gemini	5	39
	GPT-4o	11	47
	Llama	19	36
	Pixtral	11	28
Hockey Puck	Gemini	13	50
House	Claude	5	46
	Gemini	5	50
	GPT-4 Turbo	6	50
	GPT-4o	5	28
	Llama	6	47
	Pixtral	6	50
Key	Claude	10	38
	Gemini	10	50
	GPT-4 Turbo	11	50
	GPT-4o	11	32
	Llama	15	30
	Pixtral	14	46
Radio	Claude	10	33
	Gemini	5	50
	GPT-4 Turbo	17	50
	GPT-4o	16	26
	Llama	10	29
	Pixtral	19	36
String Bean	Gemini	13	49
Sun	Claude	7	46
	Gemini	5	30
	GPT-4 Turbo	7	50
	GPT-4o	9	39
	Llama	7	25
	Pixtral	9	45
Sword	Claude	6	50
	Gemini	5	50
	GPT-4 Turbo	7	50
	GPT-4o	8	50
	Llama	7	39
	Pixtral	7	46
Television	Claude	6	49
	Gemini	5	50
	GPT-4 Turbo	7	50
	GPT-4o	6	49
	Llama	6	47
	Pixtral	6	46

Continued on next page.

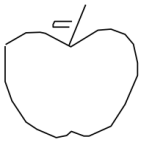
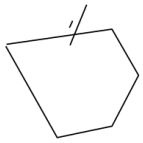
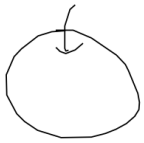
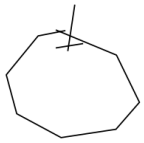
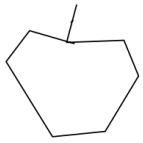
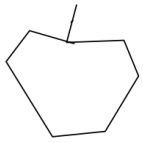
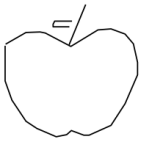
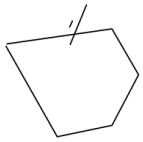
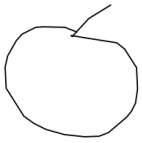
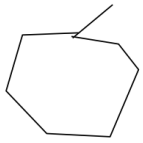
Concept	Model	$\text{TS}_{0.5,50}(c)$	Correct
Tree	Claude	7	41
	Gemini	9	49
	GPT-4 Turbo	14	50
	GPT-4o	4	48
	Llama	9	26
	Pixtral	10	26

Table 6: This table presents the teaching size (TS) for each concept in the coordinates modality. The teaching size represents the minimal number of segments required in a drawing for a learner (i.e., a large language model) to recognize the concept with a probability of at least ρ over N independent trials. A lower teaching size indicates a simpler representation of the concept that is still consistently identifiable by the model.

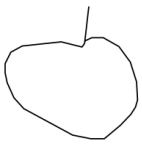
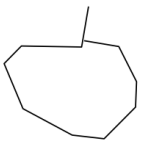
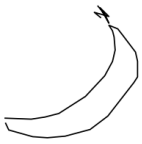
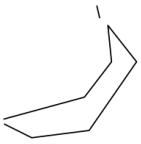
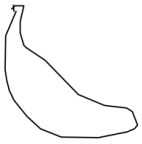
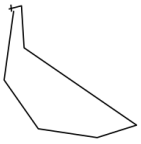
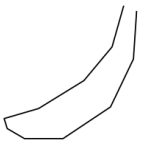
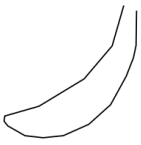
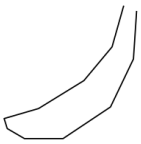
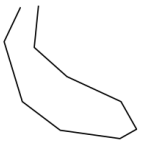
Concept	Model	$\mathrm{TS}_{0.5,50}(c)$	Correct
Apple	Claude	10	31
	GPT-4o	33	31
Banana	Gemini	13	48
Car	GPT-4 Turbo	31	50
Cat	Claude	12	41
	Gemini	19	46
	GPT-4 Turbo	20	50
	GPT-4o	13	32
	Llama	7	26
	Pixtral	21	28
Cup	Claude	15	26
Envelope	Claude	4	29
	Gemini	5	28
	GPT-4 Turbo	5	50
	GPT-4o	4	28
	Llama	4	29
Fish	Claude	11	36
	Gemini	8	48
	GPT-4 Turbo	15	32
Grass	Claude	37	27
	Gemini	17	44
House	Claude	3	26
	Gemini	6	50
	GPT-4 Turbo	5	50
	GPT-4o	5	30
	Llama	4	41
	Pixtral	5	43
Sun	Claude	13	42
	Gemini	7	50
Tree	Gemini	7	46
	GPT-4 Turbo	15	47
	GPT-4o	15	26

E Original and Simplest Drawing for each Model and Modality

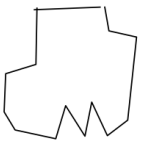
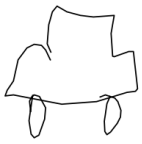
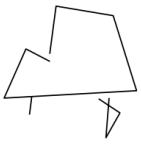
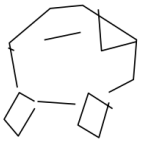
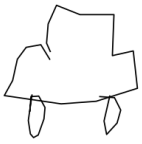
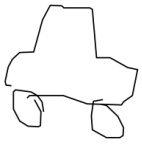
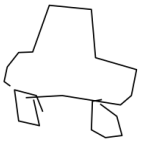
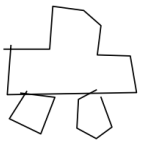
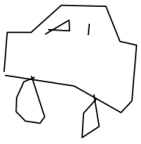
Table 7: Original and simplest drawing for each concept, modality and model. For each concept, the table includes both the original drawing and its simplified version, as processed by the Ramer–Douglas–Peucker algorithm. The original drawings are those directly sourced from the *Quick, Draw!* dataset. In contrast, the simplest drawings result from iterative simplification, which reduces the number of segments while preserving the essential characteristics of the concept. This simplified version represents the minimal form that each learner (i.e., large language model) can still recognize with a high probability (as per the definition of teaching size of this work). By comparing these drawings, we can better understand the inherent simplicity or complexity of each concept and how it translates across visual and textual representations.

Concept	Model	Original (images)	Simplified (images)	Original (coordinates)	Simplified (coordinates)
Apple	Claude				
	Gemini				
	GPT-4 Turbo				
	GPT-4o				
	Llama				

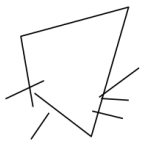
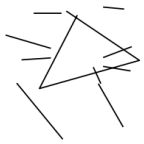
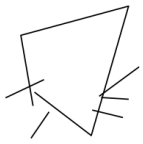
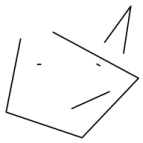
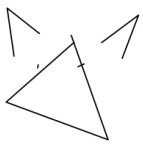
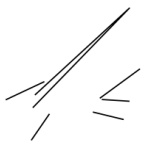
Continued on next page.

Concept	Model	Original (images)	Simplified (images)	Original (coordinates)	Simplified (coordinates)
Banana	Pixtral				
	Claude				
	Gemini				
	GPT-4 Turbo				
	GPT-4o				
	Llama				
	Pixtral				

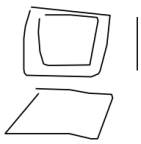
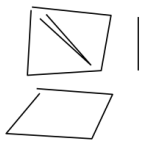
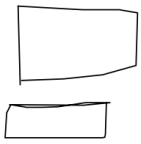
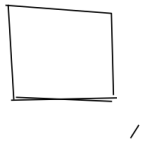
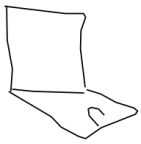
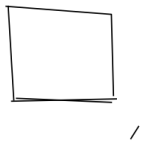
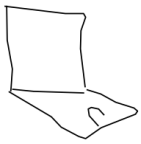
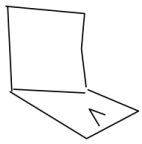
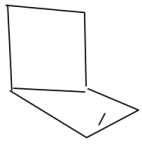
Continued on next page.

Concept	Model	Original (images)	Simplified (images)	Original (coordinates)	Simplified (coordinates)
Car	Claude				
	Gemini				
	GPT-4 Turbo				
	GPT-4o				
	Llama				
	Pixtral				

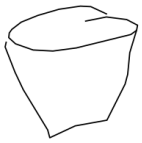
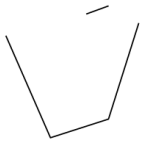
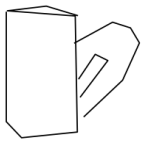
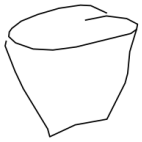
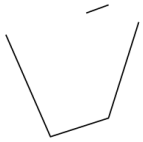
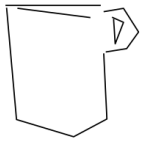
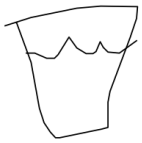
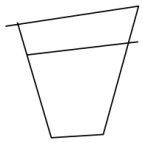
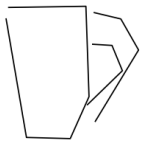
Continued on next page.

Concept	Model	Original (images)	Simplified (images)	Original (coordinates)	Simplified (coordinates)
Cat	Claude				
	Gemini				
	GPT-4 Turbo				
	GPT-4o				
	Llama				
	Pixtral				

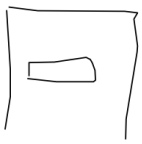
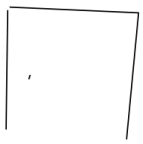
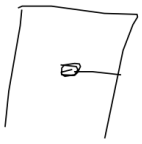
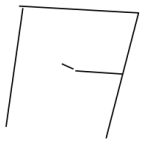
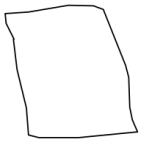
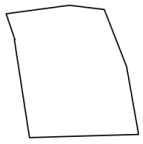
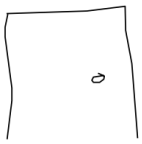
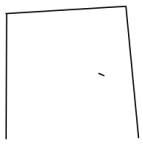
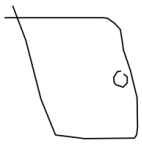
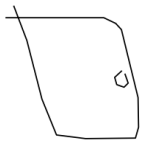
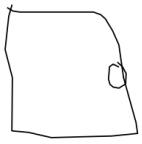
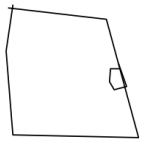
Continued on next page.

Concept	Model	Original (images)	Simplified (images)	Original (coordinates)	Simplified (coordinates)
Computer	Claude				
	Gemini				
	GPT-4 Turbo				
	GPT-4o				
	Llama				
	Pixtral				

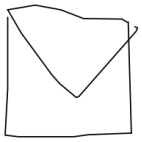
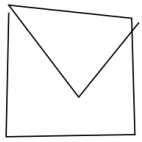
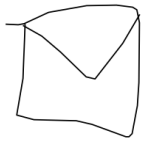
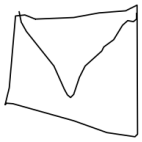
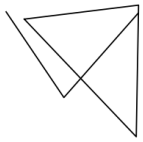
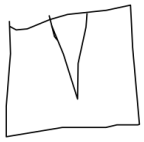
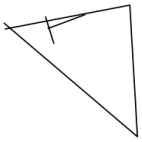
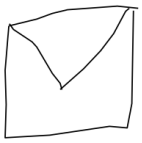
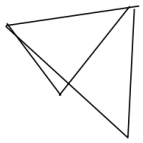
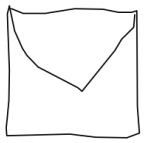
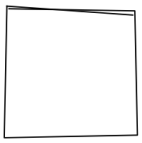
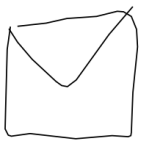
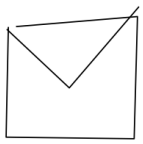
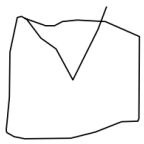
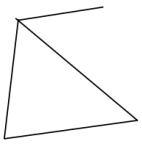
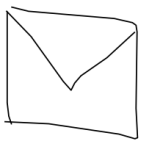
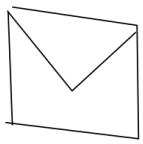
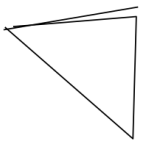
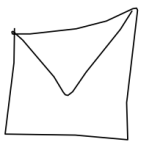
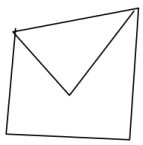
Continued on next page.

Concept	Model	Original (images)	Simplified (images)	Original (coordinates)	Simplified (coordinates)
Cup	Claude				
	Gemini				
	GPT-4 Turbo				
	GPT-4o				
	Llama				
	Pixtral				

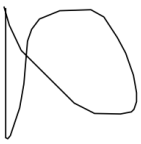
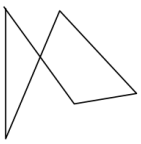
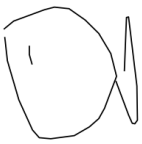
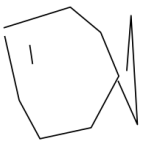
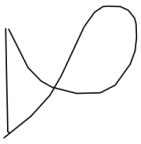
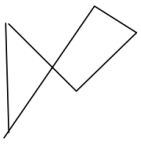
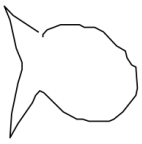
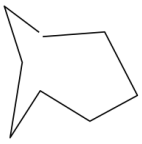
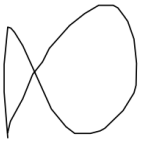
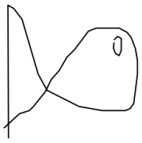
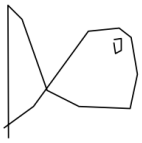
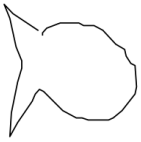
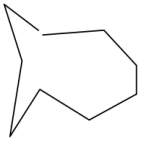
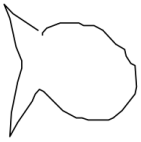
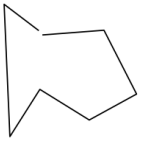
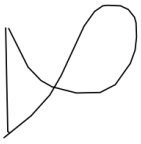
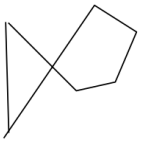
Continued on next page.

Concept	Model	Original (images)	Simplified (images)	Original (coordinates)	Simplified (coordinates)
Door	Claude				
	Gemini				
	GPT-4 Turbo				
	GPT-4o				
	Llama				
	Pixtral				

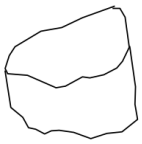
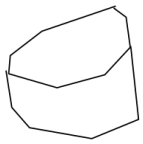
Continued on next page.

Concept	Model	Original (images)	Simplified (images)	Original (coordinates)	Simplified (coordinates)
Envelope	Claude				
	Gemini				
	GPT-4 Turbo				
	GPT-4o				
	Llama				
	Pixtral				

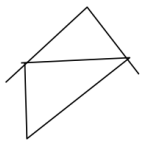
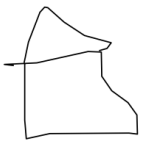
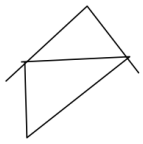
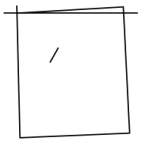
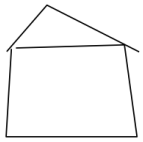
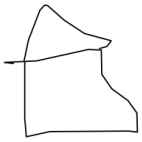
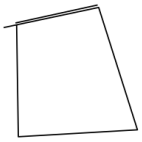
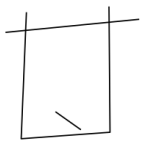
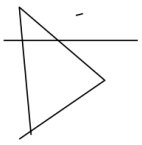
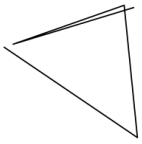
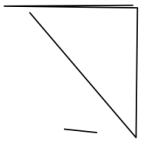
Continued on next page.

Concept	Model	Original (images)	Simplified (images)	Original (coordinates)	Simplified (coordinates)
Fish	Claude				
	Gemini				
	GPT-4 Turbo				
	GPT-4o				
	Llama				
	Pixtral				

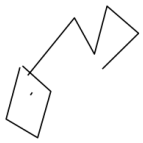
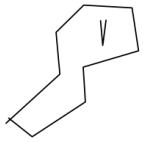
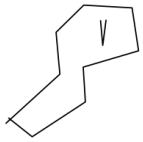
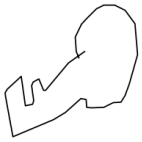
Continued on next page.

Concept	Model	Original (images)	Simplified (images)	Original (coordinates)	Simplified (coordinates)
Grass	Claude				
	Gemini				
	GPT-4o				
	Llama				
	Pixtral				
Hockey Puck	Gemini				

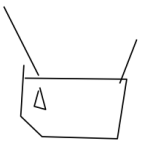
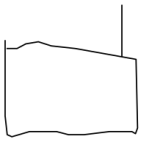
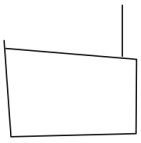
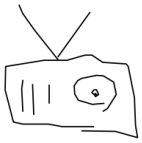
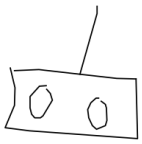
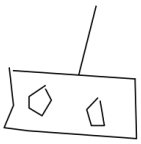
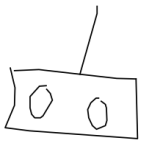
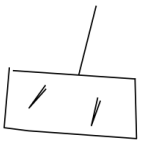
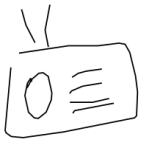
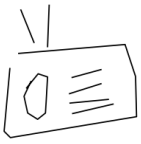
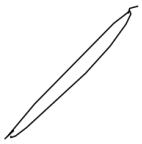
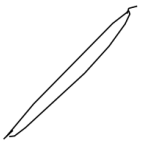
Continued on next page.

Concept	Model	Original (images)	Simplified (images)	Original (coordinates)	Simplified (coordinates)
House	Claude				
	Gemini				
	GPT-4 Turbo				
	GPT-4o				
	Llama				
	Pixtral				

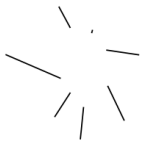
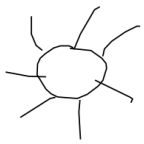
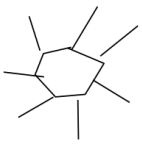
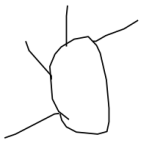
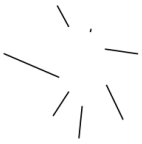
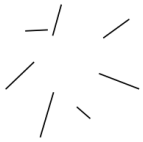
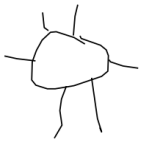
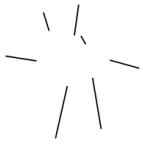
Continued on next page.

Concept	Model	Original (images)	Simplified (images)	Original (coordinates)	Simplified (coordinates)
Key	Claude				
	Gemini				
	GPT-4 Turbo				
	GPT-4o				
	Llama				
	Pixtral				

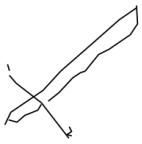
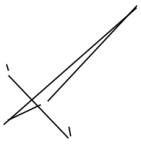
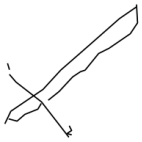
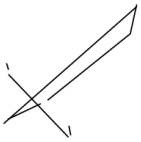
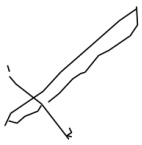
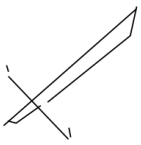
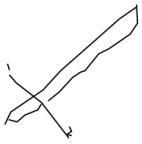
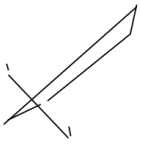
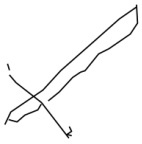
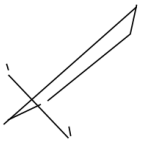
Continued on next page.

Concept	Model	Original (images)	Simplified (images)	Original (coordinates)	Simplified (coordinates)
Radio	Claude				
	Gemini				
	GPT-4 Turbo				
	GPT-4o				
	Llama				
	Pixtral				
String Bean	Gemini				

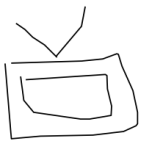
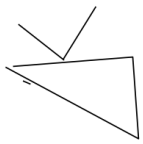
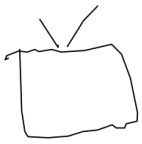
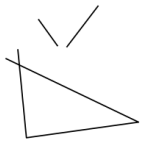
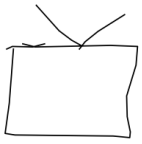
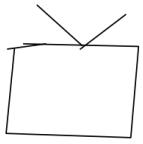
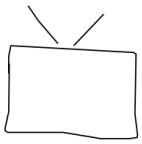
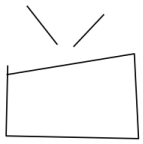
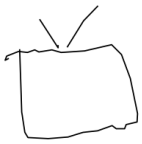
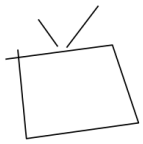
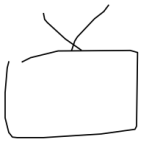
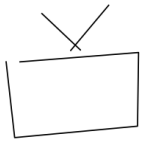
Continued on next page.

Concept	Model	Original (images)	Simplified (images)	Original (coordinates)	Simplified (coordinates)
Sun	Claude				
	Gemini				
	GPT-4 Turbo				
	GPT-4o				
	Llama				
	Pixtral				

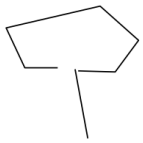
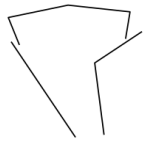
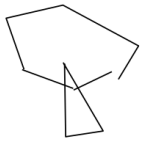
Continued on next page.

Concept	Model	Original (images)	Simplified (images)	Original (coordinates)	Simplified (coordinates)
Sword	Claude				
	Gemini				
	GPT-4 Turbo				
	GPT-4o				
	Llama				
	Pixtral				

Continued on next page.

Concept	Model	Original (images)	Simplified (images)	Original (coordinates)	Simplified (coordinates)
Television	Claude				
	Gemini				
	GPT-4 Turbo				
	GPT-4o				
	Llama				
	Pixtral				

Continued on next page.

Concept	Model	Original (images)	Simplified (images)	Original (coordinates)	Simplified (coordinates)
Tree	Claude				
	Gemini				
	GPT-4o				
	Llama				
	Pixtral				

F Confusion Tables of the Classification

Table 8: Confusion matrix for the Claude model, showing the number of times each concept is accurately predicted or misclassified in the visual-based modality (images). Each cell in the matrix represents the count of instances for a specific actual concept versus a predicted concept. The "Other" column shows the number of predictions that do not match any predefined concepts, since the model is allowed to provide open-ended answers.

Predicted concept Concept	Banana	Fish	String Bean	Sun	The Great Wall of China	Envelope	Sword	Tree	Television	Car	Hockey Puck	Grass	Cat	House	Apple	Computer	Radio	Key	Cup	Door	Other	Total
Banana	162 (32.46%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.20%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	2 (0.40%)	0 (0.00%)	334 (66.93%)	499
Fish	0 (0.00%)	266 (45.08%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.17%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	3 (0.51%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.17%)	0 (0.00%)	2 (0.34%)	317 (53.73%)	590
String Bean	58 (13.74%)	3 (0.71%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	17 (4.03%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	3 (0.71%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	3 (0.71%)	0 (0.00%)	1 (0.24%)	337 (79.86%)	422
Sun	0 (0.00%)	0 (0.00%)	0 (0.00%)	265 (61.48%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.23%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	6 (1.39%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	159 (36.89%)	431
The Great Wall of China	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.18%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	4 (0.71%)	3 (0.53%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	14 (2.50%)	539 (96.08%)	561
Envelope	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	284 (62.01%)	0 (0.00%)	1 (0.22%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	5 (1.09%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	2 (0.44%)	0 (0.00%)	166 (36.24%)	458
Sword	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	130 (27.20%)	3 (0.63%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	3 (0.63%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	2 (0.42%)	340 (71.13%)	478
Tree	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	239 (38.36%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	5 (0.80%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	3 (0.48%)	376 (60.35%)	623
Television	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	2 (0.36%)	0 (0.00%)	0 (0.00%)	385 (69.49%)	1 (0.18%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	9 (1.62%)	3 (0.54%)	1 (0.18%)	2 (0.36%)	0 (0.00%)	0 (0.00%)	1 (0.18%)	150 (27.08%)	554
Car	0 (0.00%)	12 (1.53%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.13%)	0 (0.00%)	0 (0.00%)	138 (17.60%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	57 (7.27%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	576 (73.47%)	784
Hockey Puck	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	72 (13.36%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	28 (5.19%)	4 (0.74%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	6 (1.11%)	0 (0.00%)	429 (79.59%)	539
Grass	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	4 (1.27%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	310 (98.73%)	314
Cat	0 (0.00%)	1 (0.14%)	0 (0.00%)	3 (0.42%)	0 (0.00%)	5 (0.71%)	1 (0.14%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	447 (63.05%)	4 (0.56%)	0 (0.00%)	1 (0.14%)	0 (0.00%)	0 (0.00%)	1 (0.14%)	1 (0.14%)	245 (34.56%)	709
House	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	11 (2.51%)	1 (0.23%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	282 (64.24%)	0 (0.00%)	1 (0.23%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	3 (0.68%)	141 (32.12%)	439
Apple	0 (0.00%)	2 (0.31%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.16%)	0 (0.00%)	17 (2.67%)	6 (0.94%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	28 (4.40%)	331 (51.96%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.16%)	0 (0.00%)	251 (39.40%)	637
Computer	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	9 (1.57%)	1 (0.17%)	0 (0.00%)	129 (22.51%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	6 (1.05%)	0 (0.00%)	111 (19.37%)	0 (0.00%)	0 (0.00%)	1 (0.17%)	17 (2.97%)	299 (52.18%)	573
Radio	0 (0.00%)	4 (0.62%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	5 (0.78%)	3 (0.47%)	0 (0.00%)	162 (25.27%)	3 (0.47%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	51 (7.96%)	0 (0.00%)	1 (0.16%)	95 (14.82%)	0 (0.00%)	0 (0.00%)	6 (0.94%)	311 (48.52%)	641
Key	0 (0.00%)	3 (0.41%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.14%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.14%)	4 (0.55%)	0 (0.00%)	1 (0.14%)	0 (0.00%)	117 (16.12%)	0 (0.00%)	10 (1.38%)	589 (81.13%)	726
Cup	0 (0.00%)	9 (1.54%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	31 (5.30%)	0 (0.00%)	0 (0.00%)	3 (0.51%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	7 (1.20%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	221 (37.78%)	24 (4.10%)	290 (49.57%)	585
Door	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	3 (0.78%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	84 (21.82%)	298 (77.40%)	385
Total	220	300	0	268	0	419	159	244	696	148	0	4	452	507	338	116	97	121	234	168	6457	10948

Table 9: Confusion matrix for the Claude model, showing the number of times each concept is accurately predicted or misclassified in the text-based modality (coordinates). Each cell in the matrix represents the count of instances for a specific actual concept versus a predicted concept. The "Other" column shows the number of predictions that do not match any predefined concepts, since the model is allowed to provide open-ended answers.

Predicted concept Concept	Banana	Fish	String Bean	Sun	The Great Wall of China	Envelope	Sword	Tree	Television	Car	Hockey Puck	Grass	Cat	House	Apple	Computer	Radio	Key	Cup	Door	Other	Total
Banana	0 (0.00%)	89 (17.84%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.20%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	8 (1.60%)	4 (0.80%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	397 (79.56%)	499
Fish	0 (0.00%)	109 (18.47%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.17%)	0 (0.00%)	1 (0.17%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	34 (5.76%)	25 (4.24%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	420 (71.19%)	590
String Bean	0 (0.00%)	50 (11.85%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	3 (0.71%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	4 (0.95%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	365 (86.49%)	422
Sun	0 (0.00%)	21 (4.87%)	0 (0.00%)	68 (15.78%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	3 (0.70%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	4 (0.93%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	335 (77.73%)	431
The Great Wall of China	0 (0.00%)	12 (2.14%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	3 (0.53%)	0 (0.00%)	56 (9.98%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	2 (0.36%)	0 (0.00%)	488 (86.99%)	561
Envelope	0 (0.00%)	35 (7.64%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	44 (9.61%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	115 (25.11%)	1 (0.22%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	263 (57.42%)	458
Sword	0 (0.00%)	91 (19.04%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	8 (1.67%)	0 (0.00%)	5 (1.05%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	2 (0.42%)	41 (8.58%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	331 (69.25%)	478
Tree	0 (0.00%)	269 (43.18%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	3 (0.48%)	0 (0.00%)	1 (0.16%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	54 (8.67%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.16%)	0 (0.00%)	295 (47.35%)	623
Television	0 (0.00%)	80 (14.44%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	41 (7.40%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.18%)	0 (0.00%)	0 (0.00%)	9 (1.62%)	278 (50.18%)	2 (0.36%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	18 (3.25%)	0 (0.00%)	125 (22.56%)	554
Car	0 (0.00%)	183 (23.34%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	4 (0.51%)	0 (0.00%)	6 (0.77%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	18 (2.30%)	101 (12.88%)	7 (0.89%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	3 (0.38%)	0 (0.00%)	462 (58.93%)	784
Hockey Puck	0 (0.00%)	97 (18.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	18 (3.34%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	51 (9.46%)	17 (3.15%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.19%)	0 (0.00%)	355 (65.86%)	539
Grass	0 (0.00%)	1 (0.32%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	5 (1.59%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	8 (2.55%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	300 (95.54%)	314
Cat	0 (0.00%)	190 (26.80%)	0 (0.00%)	1 (0.14%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	11 (1.55%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	6 (0.85%)	164 (23.13%)	18 (2.54%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	319 (44.99%)	709
House	0 (0.00%)	71 (16.17%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	16 (3.64%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	161 (36.67%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	191 (43.51%)	439
Apple	0 (0.00%)	151 (23.70%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	66 (10.36%)	63 (9.89%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	3 (0.47%)	0 (0.00%)	354 (55.57%)	637
Computer	0 (0.00%)	69 (12.04%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	18 (3.14%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.17%)	0 (0.00%)	1 (0.17%)	0 (0.00%)	284 (49.56%)	1 (0.17%)	1 (0.17%)	0 (0.00%)	0 (0.00%)	22 (3.84%)	1 (0.17%)	175 (30.54%)	573
Radio	0 (0.00%)	117 (18.25%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	22 (3.43%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.16%)	0 (0.00%)	1 (0.16%)	25 (3.90%)	230 (35.88%)	4 (0.62%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	9 (1.40%)	0 (0.00%)	232 (36.19%)	641
Key	0 (0.00%)	266 (36.64%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	7 (0.96%)	0 (0.00%)	3 (0.41%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	2 (0.28%)	1 (0.14%)	77 (10.61%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	13 (1.79%)	0 (0.00%)	357 (49.17%)	726
Cup	0 (0.00%)	126 (21.54%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	13 (2.22%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	122 (20.85%)	10 (1.71%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	17 (2.91%)	0 (0.00%)	297 (50.77%)	585
Door	0 (0.00%)	21 (5.45%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	5 (1.30%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	3 (0.78%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	105 (27.27%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	10 (2.60%)	0 (0.00%)	241 (62.60%)	385
Total	0	2048	0	69	0	204	0	35	0	6	0	21	223	1805	134	1	0	0	99	1	6302	10948

Table 10: Confusion matrix for the Gemini model, showing the number of times each concept is accurately predicted or misclassified in the visual-based modality (images). Each cell in the matrix represents the count of instances for a specific actual concept versus a predicted concept. The "Other" column shows the number of predictions that do not match any predefined concepts, since the model is allowed to provide open-ended answers.

Predicted concept Concept	Banana	Fish	String Bean	Sun	The Great Wall of China	Envelope	Sword	Tree	Television	Car	Hockey Puck	Grass	Cat	House	Apple	Computer	Radio	Key	Cup	Door	Other	Total
Banana	148 (29.66%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.20%)	1 (0.20%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.20%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	348 (69.74%)	499
Fish	0 (0.00%)	248 (42.03%)	0 (0.00%)	2 (0.34%)	0 (0.00%)	2 (0.34%)	1 (0.17%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.17%)	0 (0.00%)	2 (0.34%)	0 (0.00%)	1 (0.17%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	333 (56.44%)	590
String Bean	53 (12.56%)	0 (0.00%)	11 (2.61%)	0 (0.00%)	0 (0.00%)	1 (0.24%)	18 (4.27%)	0 (0.00%)	1 (0.24%)	0 (0.00%)	0 (0.00%)	1 (0.24%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.24%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	336 (79.62%)	422
Sun	0 (0.00%)	0 (0.00%)	0 (0.00%)	283 (65.66%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	148 (34.34%)	431
The Great Wall of China	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	2 (0.36%)	1 (0.18%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	9 (1.60%)	2 (0.36%)	3 (0.53%)	0 (0.00%)	1 (0.18%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	543 (96.79%)	561
Envelope	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	320 (69.87%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.22%)	0 (0.00%)	3 (0.66%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	134 (29.26%)	458
Sword	0 (0.00%)	1 (0.21%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	188 (39.33%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.21%)	0 (0.00%)	1 (0.21%)	0 (0.00%)	2 (0.42%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	285 (59.62%)	478
Tree	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.16%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	248 (39.81%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	2 (0.32%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	372 (59.71%)	623
Television	0 (0.00%)	0 (0.00%)	0 (0.00%)	3 (0.54%)	0 (0.00%)	2 (0.36%)	0 (0.00%)	0 (0.00%)	412 (74.37%)	0 (0.00%)	0 (0.00%)	1 (0.18%)	1 (0.18%)	0 (0.00%)	0 (0.00%)	12 (2.17%)	1 (0.18%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	122 (22.02%)	554
Car	0 (0.00%)	21 (2.68%)	0 (0.00%)	1 (0.13%)	0 (0.00%)	3 (0.38%)	1 (0.13%)	1 (0.13%)	0 (0.00%)	333 (42.47%)	0 (0.00%)	1 (0.13%)	0 (0.00%)	12 (1.53%)	0 (0.00%)	3 (0.38%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	408 (52.04%)	784
Hockey Puck	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	25 (4.64%)	2 (0.37%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	87 (16.14%)	1 (0.19%)	0 (0.00%)	12 (2.23%)	0 (0.00%)	16 (2.97%)	0 (0.00%)	0 (0.00%)	9 (1.67%)	0 (0.00%)	387 (71.80%)	539
Grass	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	111 (35.35%)	0 (0.00%)	1 (0.32%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	202 (64.33%)	314
Cat	0 (0.00%)	2 (0.28%)	0 (0.00%)	10 (1.41%)	0 (0.00%)	1 (0.14%)	1 (0.14%)	0 (0.00%)	1 (0.14%)	0 (0.00%)	0 (0.00%)	2 (0.28%)	529 (74.61%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.14%)	162 (22.85%)	709
House	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	13 (2.96%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	282 (64.24%)	0 (0.00%)	4 (0.91%)	0 (0.00%)	0 (0.00%)	1 (0.23%)	0 (0.00%)	139 (31.66%)	439
Apple	0 (0.00%)	0 (0.00%)	0 (0.00%)	2 (0.31%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	11 (1.73%)	1 (0.16%)	0 (0.00%)	1 (0.16%)	1 (0.16%)	5 (0.78%)	270 (42.39%)	1 (0.16%)	6 (0.94%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	339 (53.22%)	637
Computer	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.17%)	0 (0.00%)	1 (0.17%)	0 (0.00%)	0 (0.00%)	79 (13.79%)	0 (0.00%)	0 (0.00%)	2 (0.35%)	0 (0.00%)	4 (0.70%)	0 (0.00%)	277 (48.34%)	0 (0.00%)	0 (0.00%)	1 (0.17%)	1 (0.17%)	207 (36.13%)	573
Radio	0 (0.00%)	0 (0.00%)	0 (0.00%)	2 (0.31%)	0 (0.00%)	0 (0.00%)	1 (0.16%)	0 (0.00%)	81 (12.64%)	0 (0.00%)	1 (0.16%)	1 (0.16%)	2 (0.31%)	7 (1.09%)	0 (0.00%)	5 (0.78%)	336 (52.42%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	205 (31.98%)	641
Key	0 (0.00%)	5 (0.69%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	5 (0.69%)	0 (0.00%)	0 (0.00%)	4 (0.55%)	0 (0.00%)	0 (0.00%)	1 (0.14%)	0 (0.00%)	0 (0.00%)	6 (0.83%)	0 (0.00%)	228 (31.40%)	0 (0.00%)	0 (0.00%)	477 (65.70%)	726
Cup	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.17%)	0 (0.00%)	15 (2.56%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	2 (0.34%)	0 (0.00%)	6 (1.03%)	0 (0.00%)	4 (0.68%)	0 (0.00%)	0 (0.00%)	295 (50.43%)	2 (0.34%)	260 (44.44%)	585
Door	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	2 (0.52%)	0 (0.00%)	0 (0.00%)	1 (0.26%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	8 (2.08%)	0 (0.00%)	9 (2.34%)	0 (0.00%)	0 (0.00%)	7 (1.82%)	71 (18.44%)	287 (74.55%)	385
Total	201	277	11	306	0	388	219	249	586	338	88	134	536	346	270	346	343	228	313	75	5694	10948

Table 11: Confusion matrix for the Gemini model, showing the number of times each concept is accurately predicted or misclassified in the text-based modality (coordinates). Each cell in the matrix represents the count of instances for a specific actual concept versus a predicted concept. The "Other" column shows the number of predictions that do not match any predefined concepts, since the model is allowed to provide open-ended answers.

Predicted concept Concept	Banana	Fish	String Bean	Sun	The Great Wall of China	Envelope	Sword	Tree	Television	Car	Hockey Puck	Grass	Cat	House	Apple	Computer	Radio	Key	Cup	Door	Other	Total
Banana	9 (1.80%)	107 (21.44%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	383 (76.75%)	499
Fish	2 (0.34%)	172 (29.15%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	8 (1.36%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.17%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	407 (68.98%)	590
String Bean	8 (1.90%)	56 (13.27%)	0 (0.00%)	2 (0.47%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	2 (0.47%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	354 (83.89%)	422
Sun	0 (0.00%)	17 (3.94%)	0 (0.00%)	175 (40.60%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	29 (6.73%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	3 (0.70%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	207 (48.03%)	431
The Great Wall of China	3 (0.53%)	13 (2.32%)	0 (0.00%)	15 (2.67%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	40 (7.13%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	6 (1.07%)	0 (0.00%)	3 (0.53%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	481 (85.74%)	561
Envelope	3 (0.66%)	69 (15.07%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	3 (0.66%)	0 (0.00%)	2 (0.44%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	3 (0.66%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	378 (82.53%)	458
Sword	0 (0.00%)	38 (7.95%)	0 (0.00%)	1 (0.21%)	0 (0.00%)	1 (0.21%)	0 (0.00%)	43 (9.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	3 (0.63%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	392 (82.01%)	478
Tree	2 (0.32%)	26 (4.17%)	0 (0.00%)	16 (2.57%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	129 (20.71%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.16%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	449 (72.07%)	623
Television	0 (0.00%)	95 (17.15%)	0 (0.00%)	4 (0.72%)	0 (0.00%)	8 (1.44%)	0 (0.00%)	16 (2.89%)	0 (0.00%)	1 (0.18%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	83 (14.98%)	1 (0.18%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	346 (62.45%)	554
Car	0 (0.00%)	190 (24.23%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	2 (0.26%)	0 (0.00%)	22 (2.81%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	20 (2.55%)	1 (0.13%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	549 (70.03%)	784
Hockey Puck	3 (0.56%)	195 (36.18%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.19%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	3 (0.56%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	337 (62.52%)	539
Grass	0 (0.00%)	11 (3.50%)	0 (0.00%)	36 (11.46%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	4 (1.27%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	6 (1.91%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	257 (81.85%)	314
Cat	0 (0.00%)	98 (13.82%)	0 (0.00%)	57 (8.04%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	48 (6.77%)	0 (0.00%)	1 (0.14%)	0 (0.00%)	0 (0.00%)	2 (0.28%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	503 (70.94%)	709
House	0 (0.00%)	50 (11.39%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	9 (2.05%)	0 (0.00%)	20 (4.56%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	19 (4.33%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	341 (77.68%)	439
Apple	4 (0.63%)	146 (22.92%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.16%)	0 (0.00%)	7 (1.10%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	2 (0.31%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	477 (74.88%)	637
Computer	0 (0.00%)	108 (18.85%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.17%)	1 (0.17%)	16 (2.79%)	0 (0.00%)	11 (1.92%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	101 (17.63%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.17%)	0 (0.00%)	0 (0.00%)	334 (58.29%)	573
Radio	0 (0.00%)	77 (12.01%)	0 (0.00%)	3 (0.47%)	0 (0.00%)	1 (0.16%)	0 (0.00%)	34 (5.30%)	0 (0.00%)	1 (0.16%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	79 (12.32%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	446 (69.58%)	641
Key	0 (0.00%)	293 (40.36%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	6 (0.83%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.14%)	0 (0.00%)	0 (0.00%)	1 (0.14%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	425 (58.54%)	726
Cup	3 (0.51%)	149 (25.47%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	6 (1.03%)	0 (0.00%)	7 (1.20%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	4 (0.68%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	416 (71.11%)	585
Door	1 (0.26%)	20 (5.19%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	3 (0.78%)	0 (0.00%)	5 (1.30%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	4 (1.04%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	352 (91.43%)	385
Total	38	1930	0	309	0	36	1	438	0	14	0	12	2	330	2	0	0	1	1	0	7834	10948

Table 12: Confusion matrix showing for the GPT-4 Turbo the number of times each concept is accurately predicted or misclassified in the visual-based modality (images). Each cell in the matrix represents the count of instances for a specific actual concept versus a predicted concept. The "Other" column shows the number of predictions that do not match any predefined concepts, since the model is allowed to provide open-ended answers.

Predicted concept Concept	Banana	Fish	String Bean	Sun	The Great Wall of China	Envelope	Sword	Tree	Television	Car	Hockey Puck	Grass	Cat	House	Apple	Computer	Radio	Key	Cup	Door	Other	Total
Banana	78 (15.63%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.20%)	0 (0.00%)	420 (84.17%)	499
Fish	0 (0.00%)	149 (25.25%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	2 (0.34%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	3 (0.51%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	2 (0.34%)	0 (0.00%)	0 (0.00%)	434 (73.56%)	590
String Bean	15 (3.55%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.24%)	2 (0.47%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	404 (95.73%)	422
Sun	0 (0.00%)	0 (0.00%)	0 (0.00%)	174 (40.37%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	257 (59.63%)	431
The Great Wall of China	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.18%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.18%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	6 (1.07%)	553 (98.57%)	561
Envelope	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	277 (60.48%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	181 (39.52%)	458
Sword	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.21%)	0 (0.00%)	0 (0.00%)	95 (19.87%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.21%)	0 (0.00%)	1 (0.21%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	380 (79.50%)	478
Tree	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.16%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	187 (30.02%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	3 (0.48%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	16 (2.57%)	1 (0.16%)	415 (66.61%)	623
Television	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	5 (0.90%)	1 (0.18%)	0 (0.00%)	243 (43.86%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.18%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.18%)	0 (0.00%)	4 (0.72%)	4 (0.72%)	295 (53.25%)	554
Car	0 (0.00%)	15 (1.91%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	3 (0.38%)	0 (0.00%)	112 (14.29%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	14 (1.79%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.13%)	639 (81.51%)	784
Hockey Puck	0 (0.00%)	1 (0.19%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	53 (9.83%)	1 (0.19%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	9 (1.67%)	0 (0.00%)	1 (0.19%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	474 (87.94%)	539
Grass	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.32%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	313 (99.68%)	314
Cat	0 (0.00%)	0 (0.00%)	0 (0.00%)	10 (1.41%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	400 (56.42%)	1 (0.14%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	3 (0.42%)	1 (0.14%)	294 (41.47%)	709
House	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	36 (8.20%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	248 (56.49%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	155 (35.31%)	439
Apple	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.16%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	2 (0.31%)	235 (36.89%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	2 (0.31%)	0 (0.00%)	397 (62.32%)	637
Computer	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	8 (1.40%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	3 (0.52%)	0 (0.00%)	101 (17.63%)	0 (0.00%)	0 (0.00%)	4 (0.70%)	7 (1.22%)	450 (78.53%)	573
Radio	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	24 (3.74%)	0 (0.00%)	0 (0.00%)	26 (4.06%)	40 (6.24%)	0 (0.00%)	0 (0.00%)	1 (0.16%)	20 (3.12%)	0 (0.00%)	2 (0.31%)	41 (6.40%)	0 (0.00%)	3 (0.47%)	0 (0.00%)	484 (75.51%)	641
Key	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	3 (0.41%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	151 (20.80%)	1 (0.14%)	4 (0.55%)	567 (78.10%)	726
Cup	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	19 (3.25%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.17%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	77 (13.16%)	4 (0.68%)	484 (82.74%)	585
Door	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	8 (2.08%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	2 (0.52%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.26%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	2 (0.52%)	31 (8.05%)	341 (88.57%)	385
Total	93	165	0	187	0	437	99	190	270	154	0	0	402	307	235	105	42	153	113	59	7937	10,948

Table 13: Confusion matrix for the GPT-4 Turbo model, showing the number of times each concept is accurately predicted or misclassified in the text-based modality (coordinates). Each cell in the matrix represents the count of instances for a specific actual concept versus a predicted concept. The "Other" column shows the number of predictions that do not match any predefined concepts, since the model is allowed to provide open-ended answers.

Predicted concept Concept	Banana	Fish	String Bean	Sun	The Great Wall China	Envelope	Sword	Tree	Television	Car	Hockey Puck	Grass	Cat	House	Apple	Computer	Radio	Key	Cup	Door	Other	Total
Banana	0 (0.00%)	6 (1.20%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	8 (1.60%)	0 (0.00%)	2 (0.40%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	106 (21.24%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	377 (75.55%)	499
Fish	0 (0.00%)	15 (2.54%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.17%)	0 (0.00%)	14 (2.37%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	122 (20.68%)	1 (0.17%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	437 (74.07%)	590
String Bean	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	2 (0.47%)	0 (0.00%)	1 (0.24%)	0 (0.00%)	2 (0.47%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	115 (27.25%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	302 (71.56%)	422
Sun	0 (0.00%)	8 (1.86%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	28 (6.50%)	0 (0.00%)	1 (0.23%)	0 (0.00%)	0 (0.00%)	2 (0.46%)	138 (32.02%)	1 (0.23%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	253 (58.70%)	431
TheGreat Wall of China	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	7 (1.25%)	0 (0.00%)	18 (3.21%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	85 (15.15%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	451 (80.39%)	561
Envelope	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	25 (5.46%)	0 (0.00%)	2 (0.44%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	230 (50.22%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	201 (43.89%)	458
Sword	0 (0.00%)	3 (0.63%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	2 (0.42%)	0 (0.00%)	48 (10.04%)	0 (0.00%)	3 (0.63%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	152 (31.80%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.21%)	0 (0.00%)	0 (0.00%)	269 (56.28%)	478
Tree	0 (0.00%)	4 (0.64%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.16%)	0 (0.00%)	162 (26.00%)	0 (0.00%)	6 (0.96%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	173 (27.77%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	277 (44.46%)	623
Television	0 (0.00%)	3 (0.54%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	22 (3.97%)	0 (0.00%)	2 (0.36%)	0 (0.00%)	56 (10.11%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	386 (69.68%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	85 (15.34%)	554
Car	0 (0.00%)	4 (0.51%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.13%)	0 (0.00%)	64 (8.16%)	0 (0.00%)	31 (3.95%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	324 (41.33%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	360 (45.92%)	784
Hockey Puck	0 (0.00%)	11 (2.04%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	15 (2.78%)	0 (0.00%)	4 (0.74%)	0 (0.00%)	5 (0.93%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	269 (49.91%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	235 (43.60%)	539
Grass	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	30 (9.55%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	3 (0.96%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	281 (89.49%)	314
Cat	0 (0.00%)	39 (5.50%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	68 (9.59%)	0 (0.00%)	48 (6.77%)	0 (0.00%)	0 (0.00%)	16 (2.26%)	275 (38.79%)	1 (0.14%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	262 (36.95%)	709
House	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	7 (1.59%)	0 (0.00%)	3 (0.68%)	0 (0.00%)	4 (0.91%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	280 (63.78%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	145 (33.03%)	439
Apple	0 (0.00%)	11 (1.73%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	30 (4.71%)	0 (0.00%)	5 (0.78%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	308 (48.35%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	2 (0.31%)	0 (0.00%)	0 (0.00%)	281 (44.11%)	637
Computer	0 (0.00%)	1 (0.17%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	21 (3.66%)	0 (0.00%)	2 (0.35%)	0 (0.00%)	49 (8.55%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	425 (74.17%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	75 (13.09%)	573
Radio	0 (0.00%)	9 (1.40%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	4 (0.62%)	0 (0.00%)	4 (0.62%)	0 (0.00%)	107 (16.69%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	440 (68.64%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	77 (12.01%)	641
Key	0 (0.00%)	1 (0.14%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.14%)	0 (0.00%)	58 (7.99%)	0 (0.00%)	9 (1.24%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	309 (42.56%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	348 (47.93%)	726
Cup	0 (0.00%)	42 (7.18%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	7 (1.20%)	0 (0.00%)	10 (1.71%)	0 (0.00%)	17 (2.91%)	0 (0.00%)	0 (0.00%)	1 (0.17%)	313 (53.50%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	195 (33.33%)	585
Door	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	14 (3.64%)	0 (0.00%)	7 (1.82%)	0 (0.00%)	2 (0.52%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	219 (56.88%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	143 (37.14%)	385
Total	0	157	0	0	0	130	0	563	0	347	0	0	19	4672	3	0	0	3	0	0	5054	10,948

Table 14: Confusion matrix for the GPT-4o model, showing the number of times each concept is accurately predicted or misclassified in the visual-based modality (images). Each cell in the matrix represents the count of instances for a specific actual concept versus a predicted concept. The "Other" column shows the number of predictions that do not match any predefined concepts, since the model is allowed to provide open-ended answers.

Predicted concept Concept	Banana	Fish	String Bean	Sun	The Great Wall of China	Envelope	Sword	Tree	Television	Car	Hockey Puck	Grass	Cat	House	Apple	Computer	Radio	Key	Cup	Door	Other	Total
Banana	29 (5.81%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	470 (94.19%)	499
Fish	0 (0.00%)	33 (5.59%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.17%)	0 (0.00%)	2 (0.34%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	19 (3.22%)	4 (0.68%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	2 (0.34%)	529 (89.66%)	590
String Bean	7 (1.66%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	5 (1.18%)	0 (0.00%)	0 (0.00%)	410 (97.16%)	422
Sun	0 (0.00%)	0 (0.00%)	0 (0.00%)	81 (18.79%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	2 (0.46%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	348 (80.74%)	431
The Great Wall of China	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.18%)	0 (0.00%)	0 (0.00%)	4 (0.71%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	7 (1.25%)	11 (1.96%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	538 (95.90%)	561
Envelope	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	118 (25.76%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	5 (1.09%)	11 (2.40%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	324 (70.74%)	458
Sword	1 (0.21%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	32 (6.69%)	10 (2.09%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	2 (0.42%)	4 (0.84%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	429 (89.75%)	478
Tree	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	169 (27.13%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.16%)	20 (3.21%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	433 (69.50%)	623
Television	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	24 (4.33%)	0 (0.00%)	2 (0.36%)	262 (47.29%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	2 (0.36%)	26 (4.69%)	0 (0.00%)	5 (0.90%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	233 (42.06%)	554
Car	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	4 (0.51%)	0 (0.00%)	35 (4.46%)	0 (0.00%)	0 (0.00%)	14 (1.79%)	60 (7.65%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	671 (85.59%)	784
Hockey Puck	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	17 (3.15%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.19%)	21 (3.90%)	0 (0.00%)	4 (0.74%)	0 (0.00%)	0 (0.00%)	1 (0.19%)	0 (0.00%)	495 (91.84%)	539
Grass	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	16 (5.10%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	298 (94.90%)	314
Cat	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.14%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	403 (56.84%)	1 (0.14%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	304 (42.88%)	709
House	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	7 (1.59%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.23%)	240 (54.67%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.23%)	190 (43.28%)	439
Apple	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	4 (0.63%)	0 (0.00%)	1 (0.16%)	2 (0.31%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	3 (0.47%)	25 (3.92%)	194 (30.46%)	4 (0.63%)	0 (0.00%)	0 (0.00%)	3 (0.47%)	0 (0.00%)	401 (62.95%)	637
Computer	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	24 (4.19%)	0 (0.00%)	1 (0.17%)	43 (7.50%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	2 (0.35%)	48 (8.38%)	0 (0.00%)	70 (12.22%)	0 (0.00%)	2 (0.35%)	0 (0.00%)	4 (0.70%)	379 (66.14%)	573
Radio	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	53 (8.27%)	0 (0.00%)	0 (0.00%)	71 (11.08%)	1 (0.16%)	0 (0.00%)	0 (0.00%)	17 (2.65%)	76 (11.86%)	0 (0.00%)	14 (2.18%)	12 (1.87%)	0 (0.00%)	0 (0.00%)	1 (0.16%)	396 (61.78%)	641
Key	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	3 (0.41%)	0 (0.00%)	4 (0.55%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	3 (0.41%)	15 (2.07%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	33 (4.55%)	0 (0.00%)	2 (0.28%)	666 (91.74%)	726
Cup	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	26 (4.44%)	0 (0.00%)	2 (0.34%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	22 (3.76%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	45 (7.69%)	3 (0.51%)	487 (83.25%)	585
Door	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	5 (1.30%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	21 (5.45%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	4 (1.04%)	0 (0.00%)	13 (3.38%)	342 (88.83%)	385
Total	37	33	0	81	1	282	32	200	378	36	0	16	482	605	194	97	12	44	49	26	8343	10948

Table 15: Confusion matrix for the GPT-4o model, showing the number of times each concept is accurately predicted or misclassified in the text-based modality (coordinates). Each cell in the matrix represents the count of instances for a specific actual concept versus a predicted concept. The "Other" column shows the number of predictions that do not match any predefined concepts, since the model is allowed to provide open-ended answers.

Predicted concept Concept	Banana	Fish	String Bean	Sun	The Great Wall of China	Envelope	Sword	Tree	Television	Car	Hockey Puck	Grass	Cat	House	Apple	Computer	Radio	Key	Cup	Door	Other	Total
Banana	0 (0.00%)	2 (0.40%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	4 (0.80%)	0 (0.00%)	5 (1.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.20%)	23 (4.61%)	2 (0.40%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	462 (92.59%)	499
Fish	0 (0.00%)	1 (0.17%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	9 (1.53%)	0 (0.00%)	5 (0.85%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	15 (2.54%)	41 (6.95%)	7 (1.19%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	512 (86.78%)	590
String Bean	0 (0.00%)	4 (0.95%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	2 (0.47%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.24%)	0 (0.00%)	0 (0.00%)	2 (0.47%)	23 (5.45%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	390 (92.42%)	422
Sun	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	12 (2.78%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	27 (6.26%)	9 (2.09%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	383 (88.86%)	431
The Great Wall of China	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	10 (1.78%)	0 (0.00%)	3 (0.53%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	2 (0.36%)	92 (16.40%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	454 (80.93%)	561
Envelope	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	91 (19.87%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	115 (25.11%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	252 (55.02%)	458
Sword	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.21%)	0 (0.00%)	14 (2.93%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	77 (16.11%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	2 (0.42%)	0 (0.00%)	384 (80.33%)	478
Tree	0 (0.00%)	1 (0.16%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	27 (4.33%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	60 (9.63%)	104 (16.69%)	1 (0.16%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	430 (69.02%)	623
Television	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	74 (13.36%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	6 (1.08%)	0 (0.00%)	0 (0.00%)	3 (0.54%)	270 (48.74%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	201 (36.28%)	554
Car	0 (0.00%)	1 (0.13%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	56 (7.16%)	162 (20.72%)	1 (0.13%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	562 (71.87%)	782
Hockey Puck	0 (0.00%)	11 (2.04%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	32 (5.95%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.19%)	165 (30.67%)	5 (0.93%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	324 (60.22%)	538
Grass	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	14 (4.46%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	300 (95.54%)	314
Cat	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	35 (4.94%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	76 (10.72%)	30 (4.23%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	568 (80.11%)	709
House	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	21 (4.78%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.23%)	140 (31.89%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	277 (63.10%)	439
Apple	0 (0.00%)	2 (0.31%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	5 (0.78%)	0 (0.00%)	4 (0.63%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	2 (0.31%)	162 (25.43%)	10 (1.57%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	452 (70.96%)	637
Computer	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	45 (7.85%)	0 (0.00%)	1 (0.17%)	0 (0.00%)	7 (1.22%)	0 (0.00%)	0 (0.00%)	1 (0.17%)	265 (46.25%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	254 (44.33%)	573
Radio	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	27 (4.21%)	0 (0.00%)	1 (0.16%)	0 (0.00%)	4 (0.62%)	0 (0.00%)	0 (0.00%)	8 (1.25%)	310 (48.36%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	291 (45.40%)	641
Key	0 (0.00%)	11 (1.52%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	2 (0.28%)	0 (0.00%)	14 (1.93%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	38 (5.23%)	156 (21.49%)	2 (0.28%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.14%)	0 (0.00%)	502 (69.15%)	726
Cup	0 (0.00%)	5 (0.85%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	28 (4.79%)	0 (0.00%)	5 (0.85%)	0 (0.00%)	1 (0.17%)	0 (0.00%)	0 (0.00%)	7 (1.20%)	189 (32.31%)	4 (0.68%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	346 (59.15%)	585
Door	0 (0.00%)	1 (0.26%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	40 (10.39%)	0 (0.00%)	1 (0.26%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	161 (41.82%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	182 (47.27%)	385
Total	0	39	0	0	0	391	0	141	0	19	0	0	300	2494	32	0	0	0	3	0	7526	10945

Table 16: Confusion matrix for the Llama model, showing the number of times each concept is accurately predicted or misclassified in the visual-based modality (images). Each cell in the matrix represents the count of instances for a specific actual concept versus a predicted concept. The "Other" column shows the number of predictions that do not match any predefined concepts, since the model is allowed to provide open-ended answers.

Predicted concept Concept	Banana	Fish	String Bean	Sun	The Great Wall of China	Envelope	Sword	Tree	Television	Car	Hockey Puck	Grass	Cat	House	Apple	Computer	Radio	Key	Cup	Door	Other	Total
Banana	78 (15.63%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	3 (0.60%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	6 (1.20%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	412 (82.57%)	499
Fish	0 (0.00%)	266 (45.08%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	6 (1.02%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	4 (0.68%)	5 (0.85%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	309 (52.37%)	590
String Bean	9 (2.13%)	0 (0.00%)	0 (0.00%)	2 (0.47%)	0 (0.00%)	0 (0.00%)	22 (5.21%)	0 (0.00%)	0 (0.00%)	4 (0.95%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	4 (0.95%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	381 (90.28%)	422
Sun	0 (0.00%)	0 (0.00%)	0 (0.00%)	197 (45.71%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	4 (0.93%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	230 (53.36%)	431
The Great Wall of China	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	7 (1.25%)	0 (0.00%)	2 (0.36%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	11 (1.96%)	13 (2.32%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	2 (0.36%)	4 (0.71%)	522 (93.05%)	561
Envelope	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	279 (60.92%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	179 (39.08%)	458
Sword	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	96 (20.08%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	10 (2.09%)	0 (0.00%)	2 (0.42%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	370 (77.41%)	478
Tree	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	245 (39.33%)	0 (0.00%)	1 (0.16%)	0 (0.00%)	0 (0.00%)	1 (0.16%)	5 (0.80%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	371 (59.55%)	623
Television	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	4 (0.72%)	0 (0.00%)	0 (0.00%)	369 (66.61%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	7 (1.26%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	174 (31.41%)	554
Car	0 (0.00%)	20 (2.55%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.13%)	1 (0.13%)	0 (0.00%)	61 (7.78%)	0 (0.00%)	0 (0.00%)	1 (0.13%)	77 (9.82%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	623 (79.46%)	784
Hockey Puck	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	76 (14.10%)	1 (0.19%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	2 (0.37%)	0 (0.00%)	0 (0.00%)	17 (3.15%)	3 (0.56%)	11 (2.04%)	0 (0.00%)	0 (0.00%)	4 (0.74%)	0 (0.00%)	425 (78.85%)	539
Grass	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	6 (1.91%)	0 (0.00%)	1 (0.32%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	307 (97.77%)	314
Cat	0 (0.00%)	1 (0.14%)	0 (0.00%)	5 (0.71%)	0 (0.00%)	1 (0.14%)	0 (0.00%)	1 (0.14%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	484 (68.27%)	5 (0.71%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.14%)	0 (0.00%)	211 (29.76%)	709
House	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	27 (6.15%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	266 (60.59%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	146 (33.26%)	439
Apple	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	3 (0.47%)	0 (0.00%)	0 (0.00%)	5 (0.78%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	12 (1.88%)	333 (52.28%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	284 (44.58%)	637
Computer	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	10 (1.75%)	0 (0.00%)	0 (0.00%)	92 (16.06%)	3 (0.52%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	15 (2.62%)	0 (0.00%)	80 (13.96%)	0 (0.00%)	0 (0.00%)	4 (0.70%)	6 (1.05%)	363 (63.35%)	573
Radio	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	14 (2.18%)	0 (0.00%)	0 (0.00%)	114 (17.78%)	4 (0.62%)	0 (0.00%)	0 (0.00%)	2 (0.31%)	67 (10.45%)	0 (0.00%)	1 (0.16%)	87 (13.57%)	0 (0.00%)	0 (0.00%)	1 (0.16%)	351 (54.76%)	641
Key	0 (0.00%)	6 (0.83%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	3 (0.41%)	1 (0.14%)	0 (0.00%)	2 (0.28%)	1 (0.14%)	0 (0.00%)	0 (0.00%)	5 (0.69%)	5 (0.69%)	4 (0.55%)	0 (0.00%)	0 (0.00%)	85 (11.71%)	0 (0.00%)	1 (0.14%)	613 (84.44%)	726
Cup	0 (0.00%)	1 (0.17%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	28 (4.79%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	7 (1.20%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	217 (37.09%)	5 (0.85%)	327 (55.90%)	585
Door	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	48 (12.47%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.26%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	2 (0.52%)	12 (3.12%)	322 (83.64%)	385
Total	87	294	0	204	0	509	121	249	582	74	2	6	514	521	340	94	87	85	230	29	6920	10948

Table 17: Confusion matrix for the Llama model, showing the number of times each concept is accurately predicted or misclassified in the text-based modality (coordinates). Each cell in the matrix represents the count of instances for a specific actual concept versus a predicted concept. The "Other" column shows the number of predictions that do not match any predefined concepts, since the model is allowed to provide open-ended answers.

Predicted concept Concept	Banana	Fish	String Bean	Sun	The Great Wall of China	Envelope	Sword	Tree	Television	Car	Hockey Puck	Grass	Cat	House	Apple	Computer	Radio	Key	Cup	Door	Other	Total
Banana	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	4 (0.80%)	0 (0.00%)	1 (0.20%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	217 (43.49%)	34 (6.81%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	243 (48.70%)	499
Fish	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	4 (0.68%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	376 (63.73%)	54 (9.15%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	156 (26.44%)	590
String Bean	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	9 (2.13%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	157 (37.20%)	25 (5.92%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	231 (54.74%)	422
Sun	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	291 (67.52%)	135 (31.32%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	5 (1.16%)	431
The Great Wall of China	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	8 (1.43%)	0 (0.00%)	32 (5.70%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	329 (58.65%)	74 (13.19%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	118 (21.03%)	561
Envelope	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	19 (4.15%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	215 (46.94%)	70 (15.28%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	154 (33.62%)	458
Sword	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	2 (0.42%)	0 (0.00%)	1 (0.21%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	329 (68.83%)	67 (14.02%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	79 (16.53%)	478
Tree	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	5 (0.80%)	0 (0.00%)	1 (0.16%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	466 (74.80%)	75 (12.04%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	76 (12.20%)	623
Television	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	11 (1.99%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	264 (47.65%)	256 (46.21%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	23 (4.15%)	554
Car	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.13%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	592 (75.51%)	141 (17.98%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	50 (6.38%)	784
Hockey Puck	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	11 (2.04%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	283 (52.50%)	107 (19.85%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	138 (25.60%)	539
Grass	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	2 (0.64%)	0 (0.00%)	19 (6.05%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	172 (54.78%)	10 (3.18%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	111 (35.35%)	314
Cat	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	553 (78.00%)	141 (19.89%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	15 (2.12%)	709
House	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	6 (1.37%)	0 (0.00%)	1 (0.23%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	204 (46.47%)	141 (32.12%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	87 (19.82%)	439
Apple	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	7 (1.10%)	0 (0.00%)	1 (0.16%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	341 (53.53%)	145 (22.76%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	143 (22.45%)	637
Computer	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.17%)	0 (0.00%)	2 (0.35%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	230 (40.14%)	291 (50.79%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	49 (8.55%)	573
Radio	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.16%)	0 (0.00%)	1 (0.16%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	362 (56.47%)	234 (36.51%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	43 (6.71%)	641
Key	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	9 (1.24%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	503 (69.28%)	91 (12.53%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	123 (16.94%)	726
Cup	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	5 (0.85%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	344 (58.80%)	109 (18.63%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	127 (21.71%)	585
Door	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	10 (2.60%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	192 (49.87%)	101 (26.23%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	82 (21.30%)	385
Total	0	0	0	0	0	115	0	59	0	0	0	0	6420	2301	0	0	0	0	0	0	2053	10948

Table 18: Confusion matrix for the Pixtral model, showing the number of times each concept is accurately predicted or misclassified in the visual-based modality (images). Each cell in the matrix represents the count of instances for a specific actual concept versus a predicted concept. The "Other" column shows the number of predictions that do not match any predefined concepts, since the model is allowed to provide open-ended answers.

Predicted concept Concept	Banana	Fish	String Bean	Sun	The Great Wall of China	Envelope	Sword	Tree	Television	Car	Hockey Puck	Grass	Cat	House	Apple	Computer	Radio	Key	Cup	Door	Other	Total
Banana	102 (20.44%)	3 (0.60%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	5 (1.00%)	0 (0.00%)	1 (0.20%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	388 (77.76%)	499
Fish	0 (0.00%)	171 (28.98%)	0 (0.00%)	1 (0.17%)	0 (0.00%)	0 (0.00%)	2 (0.34%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.17%)	14 (2.37%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	3 (0.51%)	1 (0.17%)	1 (0.17%)	396 (67.12%)	590
String Bean	23 (5.45%)	7 (1.66%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	4 (0.95%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	13 (3.08%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	2 (0.47%)	0 (0.00%)	0 (0.00%)	373 (88.39%)	422
Sun	0 (0.00%)	0 (0.00%)	0 (0.00%)	119 (27.61%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.23%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	26 (6.03%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	285 (66.13%)	431
The Great Wall of China	0 (0.00%)	1 (0.18%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	3 (0.53%)	0 (0.00%)	5 (0.89%)	0 (0.00%)	2 (0.36%)	0 (0.00%)	1 (0.18%)	3 (0.53%)	12 (2.14%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	2 (0.36%)	0 (0.00%)	1 (0.18%)	531 (94.65%)	561
Envelope	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	227 (49.56%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	6 (1.31%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	225 (49.13%)	458
Sword	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	65 (13.60%)	9 (1.88%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	11 (2.30%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.21%)	392 (82.01%)	478
Tree	0 (0.00%)	1 (0.16%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	124 (19.90%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	8 (1.28%)	15 (2.41%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	5 (0.80%)	0 (0.00%)	0 (0.00%)	470 (75.44%)	623
Television	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	6 (1.08%)	0 (0.00%)	3 (0.54%)	293 (52.89%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	25 (4.51%)	0 (0.00%)	8 (1.44%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	219 (39.53%)	554
Car	0 (0.00%)	3 (0.38%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	13 (1.66%)	0 (0.00%)	30 (3.83%)	0 (0.00%)	0 (0.00%)	22 (2.81%)	113 (14.41%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	2 (0.26%)	0 (0.00%)	0 (0.00%)	601 (76.66%)	784
Hockey Puck	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	24 (4.45%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	5 (0.93%)	1 (0.19%)	0 (0.00%)	0 (0.00%)	29 (5.38%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	3 (0.56%)	4 (0.74%)	0 (0.00%)	473 (87.76%)	539
Grass	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	4 (1.27%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	11 (3.50%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	299 (95.22%)	314
Cat	0 (0.00%)	1 (0.14%)	0 (0.00%)	2 (0.28%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	8 (1.13%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	417 (58.82%)	11 (1.55%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	270 (38.08%)	709
House	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	12 (2.73%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	290 (66.06%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	137 (31.21%)	439
Apple	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	3 (0.47%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	48 (7.54%)	255 (40.03%)	0 (0.00%)	0 (0.00%)	1 (0.16%)	0 (0.00%)	0 (0.00%)	330 (51.81%)	637
Computer	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	5 (0.87%)	0 (0.00%)	3 (0.52%)	81 (14.14%)	2 (0.35%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	24 (4.19%)	0 (0.00%)	97 (16.93%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.17%)	360 (62.83%)	573
Radio	0 (0.00%)	2 (0.31%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	3 (0.47%)	0 (0.00%)	0 (0.00%)	41 (6.40%)	10 (1.56%)	0 (0.00%)	0 (0.00%)	2 (0.31%)	41 (6.40%)	0 (0.00%)	1 (0.16%)	21 (3.28%)	1 (0.16%)	1 (0.16%)	2 (0.31%)	516 (80.50%)	641
Key	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.14%)	6 (0.83%)	1 (0.14%)	0 (0.00%)	1 (0.14%)	0 (0.00%)	0 (0.00%)	1 (0.14%)	20 (2.75%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	89 (12.26%)	0 (0.00%)	1 (0.14%)	606 (83.47%)	726
Cup	0 (0.00%)	1 (0.17%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	11 (1.88%)	0 (0.00%)	1 (0.17%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.17%)	33 (5.64%)	0 (0.00%)	1 (0.17%)	0 (0.00%)	7 (1.20%)	89 (15.21%)	0 (0.00%)	441 (75.38%)	585
Door	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.26%)	2 (0.52%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	23 (5.97%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.26%)	14 (3.64%)	344 (89.35%)	385
Total	125	190	0	122	0	292	77	176	417	50	1	12	455	759	255	108	21	115	96	21	7656	10948

Table 19: Confusion matrix for the Pixtral model, showing the number of times each concept is accurately predicted or misclassified in the text-based modality (coordinates). Each cell in the matrix represents the count of instances for a specific actual concept versus a predicted concept. The "Other" column shows the number of predictions that do not match any predefined concepts, since the model is allowed to provide open-ended answers.

Predicted concept Concept	Banana	Fish	String Bean	Sun	The Great Wall of China	Envelope	Sword	Tree	Television	Car	Hockey Puck	Grass	Cat	House	Apple	Computer	Radio	Key	Cup	Door	Other	Total
Banana	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	11 (2.20%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	18 (3.61%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	470 (94.19%)	499
Fish	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	9 (1.53%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	8 (1.36%)	82 (13.90%)	3 (0.51%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	488 (82.71%)	590
String Bean	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	18 (4.27%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.24%)	37 (8.77%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	366 (86.73%)	422
Sun	0 (0.00%)	0 (0.00%)	0 (0.00%)	3 (0.70%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	30 (6.96%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	34 (7.89%)	68 (15.78%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	296 (68.68%)	431
The Great Wall of China	0 (0.00%)	1 (0.18%)	0 (0.00%)	5 (0.89%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	57 (10.16%)	0 (0.00%)	1 (0.18%)	0 (0.00%)	0 (0.00%)	10 (1.78%)	76 (13.55%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	411 (73.26%)	561
Envelope	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	3 (0.66%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	61 (13.32%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	394 (86.03%)	458
Sword	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	25 (5.23%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	6 (1.26%)	123 (25.73%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	324 (67.78%)	478
Tree	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	27 (4.33%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	20 (3.21%)	131 (21.03%)	1 (0.16%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	444 (71.27%)	623
Television	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	12 (2.17%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	21 (3.79%)	316 (57.04%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	205 (37.00%)	554
Car	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	18 (2.30%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	105 (13.39%)	212 (27.04%)	2 (0.26%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	447 (57.02%)	784
Hockey Puck	0 (0.00%)	1 (0.19%)	0 (0.00%)	1 (0.19%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	5 (0.93%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	4 (0.74%)	111 (20.59%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	417 (77.37%)	539
Grass	0 (0.00%)	0 (0.00%)	0 (0.00%)	5 (1.59%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	26 (8.28%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.32%)	2 (0.64%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	280 (89.17%)	314
Cat	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	15 (2.12%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	130 (18.34%)	177 (24.96%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	387 (54.58%)	709
House	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.23%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	8 (1.82%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	6 (1.37%)	174 (39.64%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	250 (56.95%)	439
Apple	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	5 (0.78%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	11 (1.73%)	228 (35.79%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	393 (61.70%)	637
Computer	0 (0.00%)	1 (0.17%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	13 (2.27%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	14 (2.44%)	309 (53.93%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	236 (41.19%)	573
Radio	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	10 (1.56%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	49 (7.64%)	336 (52.42%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	246 (38.38%)	641
Key	0 (0.00%)	2 (0.28%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	14 (1.93%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	38 (5.23%)	164 (22.59%)	4 (0.55%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	504 (69.42%)	726
Cup	0 (0.00%)	1 (0.17%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	9 (1.54%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	8 (1.37%)	204 (34.87%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	363 (62.05%)	585
Door	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	7 (1.82%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.26%)	137 (35.58%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	240 (62.34%)	385
Total	0	6	0	15	0	0	0	322	0	1	0	0	467	2966	10	0	0	0	0	0	7161	10948