LLM for Robots: Object Classification Study

March 12, 2025

Abstract

This study evaluates the performance of vision-based Large Language Models (LLMs) in classifying objects through a robot camera. Five models are tested on two different image sets to determine their classification accuracy. Results indicate that larger models perform better, classification is more reliable when objects are placed in isolation rather than held by the robot, and that classes should be more distinct.

1 Introduction

Advancements in vision-based LLMs have enabled robots to interpret visual input and make intelligent decisions. This study focuses on classifying objects captured by a robot's camera into predefined categories.

2 Classification Categories

Objects are classified into one of the following four categories:

- **Personal items**
- **Tools/Electronics**
- **Trash**
- **Unclear**

3 Dataset

Two sets of images were used in this study:

- **dataset vlm :** 45 images of various objects captured from the robot's camera while being held by its arm.
- **dataset table :** 38 images of objects placed on a desk, digitally altered to match the robot camera quality.

4 Classification Logic

Objects are classified based on keyword matching:

1. Identify relevant words from a predefined list.
2. Assign the object to the category with the highest number of word matches.
3. If there is a tie or no matches, classify as *unclear*.

5 User Data and Error Calculation

The subjective user classification serves as the ground truth. An error is recorded if the model's classification does not match the user's classification.

6 Prompt Used

A standardized prompt was applied across all models:

> *"You are a robot. Based on what you see, you must identify the object. You might see your robotic arm, but ignore it; focus on the object. After this: if the object should go in the personal items box (number 1), the electronics and tools box (number 2), or if the object is trash and should be thrown in the trash can (number 3). Be concise, identify the object, and you must categorize it in one of the three categories at all costs. Do not mention what it is not, only what it is and its category."*

7 Models Evaluated

The following models, available on Ollama, were tested:

- **LLaVA (7B)**
- **LLaVA-LLaMA3 (8B)**
- **LLaMA3.2-Vision (11B)**
- **LLaVA:13B (13B)**
- **LLaVA:34B (34B)**

All models were evaluated with a temperature of 0 to ensure reproducibility.

8 Results

8.1 Classification

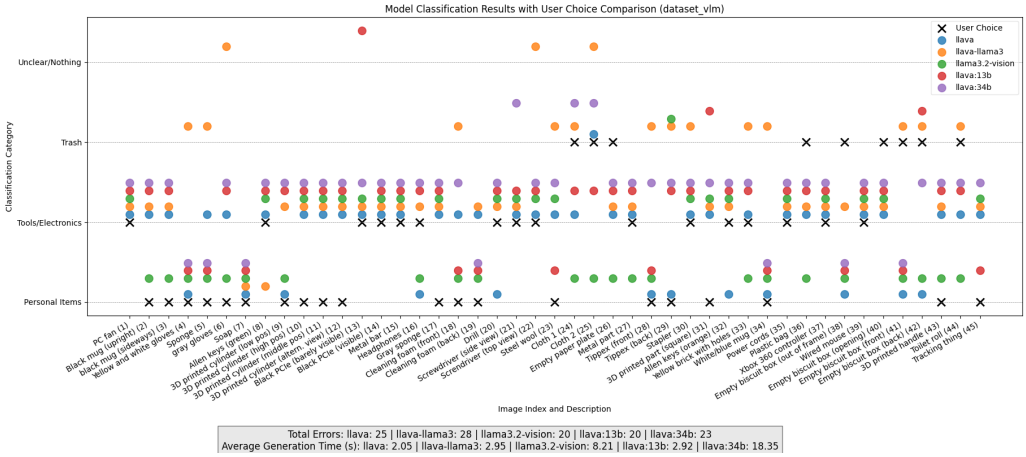

Figure 1: Classification Error Across Different Models (robot arm)

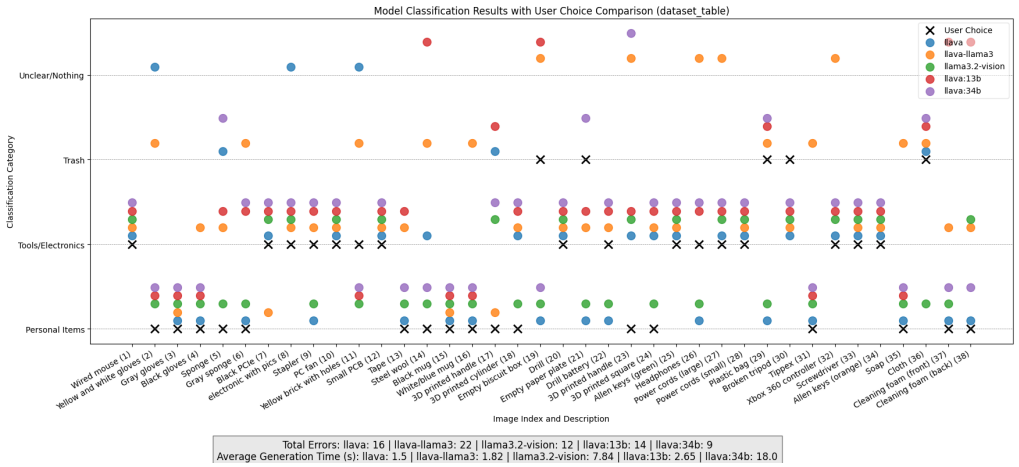

Figure 2: Classification Error Across Different Models (table)

- **Dataset vlm: Robot Holding the Object (Figure 1)**
  Model have a tendency of classifying images as Tools/Electronics, perhaps because of the robot arm present in every image. They struggle with classification accuracy, with high error rates across all models.
- **Dataset table: Object on the Desk (Figure 2)**
  LLMs perform significantly better when the object is placed in isolation on a desk.

8.2 Performance

Model	Error (robot arm)	Error (table)	Average Generation Time (s)
LLaVA (7B)	25	16	1.78
LLaVA-LLaMA3 (8B)	28	22	2.38
LLaMA3.2-Vision (11B)	20	12	8.03
LLaVA:13B (13B)	20	14	2.79
LLaVA:34B (34B)	23	9	18.18

Table 1: Performance of each model (lower is better)

Generally, when model size increases, generation time and performance increases too. However, this is not always the case. For example with the robot arm image set, llava:13b performs better than llava:34b while being around 6 times faster.

9 Conclusion

Results indicate that classification accuracy improves with larger model sizes. However, object identification remains challenging when the robot is holding the object. Performance is significantly better when objects are placed on a desk in isolation.

Moreover, the classification should be improved with better classes, due to some objects fitting multiple categories at the same time. For example, here's llama3.2-vision's response for the dirty cloth image:

> *"The object is a white cloth with red stains on it, which should be placed in the personal items box (number 1)."*

While it did correctly identify the item, it classified it as a personal item. It is not wrong depending on the user's preference, but in this case the bigger model (llava:34b) had better performance:

> *"The object appears to be a piece of cloth or paper with some stains on it. It should go in the trash can (number 3)."*