Illusory VQA: Benchmarking and Enhancing Multimodal Models on Visual Illusions

Supplementary Material

8. Task Examples

In Figures 4, 5, you find some examples of our task.

9. Dataset Examples

In Figures 6, 7, 8, and 9, you can find examples from each dataset.

10. Human Evaluation

For the human evaluation study, we present annotators with illusory images along with choices for three datasets: IllusionFashionMNIST, IllusionMNIST, and IllusionAnimals. Annotators are first asked to choose the label they suppose to be true from a set of given labels. After selecting the label, they are then shown the true label and asked to determine whether the true label is present in the image by answering with "yes" or "no." For the IllusionChar dataset, we provide images and ask annotators to type the character sequence they see. To ensure the validity of our dataset, we presented participants with the true sequence of characters in the image and asked them to confirm if they could perceive it. The results of this human evaluation are presented in Tables 5 and 6.

11. Datasets' Pie Charts

The pie charts offer an overview of the proportion of each class present in the datasets. The datasets included are IllusionMNIST, IllusionFashionMNIST, and IllusionAnimals, each analyzed separately for their training and testing splits. For the IllusionMNIST dataset, see the train split in Figure 10 and the test split in Figure 11. The IllusionFashionMNIST dataset is shown in Figure 12 for the train split and Figure 13 for the test split. Finally, the IllusionAnimals dataset splits are depicted in Figure 14 for the train split and Figure 15 for the test split.

12. Results Visualization

You can view the visualizations of the results for both zero-shot and fine-tuned models below (Figures 16, 17, 18, 19 display zero-shot results and figures 20, 21, 22, 23 display fine-tuned version). You can also find a comparison of the F1 scores for zero-shot and fine-tuned models in the figures 24, 25, 26.

13. Confusion Matrices

In this section, we present the confusion matrices for the responses generated by GPT-4o and the overall human evaluations across three different datasets: IllusionAnimals, IllusionMNIST, and IllusionFashionMNIST.

13.1. Confusion Matrices for GPT-4o

Figures 27, 28, 29, 30, 31, 32, 33, 34, 35 display the confusion matrices for the answers provided by GPT-4o on each part (Raw, Illusion, Filtered) of each dataset. We observe that GPT-4o demonstrates reasonably good performance on raw images across all classes. However, when applying illusions to the images, the values on the main diagonal of the confusion matrix decrease, indicating that it is challenging for GPT-4o to detect illusions. Conversely, after applying our filter to the illusory images, the values on the main diagonal of the confusion matrix increase, highlighting the effectiveness of our method in detecting illusions in images. In the confusion matrices, it is evident that for nearly all classes, the performance of GPT-4o significantly decreases after applying the illusion. However, after applying our filter to the images, we observe an improvement in the performance across almost all classes. This demonstrates the effectiveness of our method, despite its simplicity.

13.2. Confusion Matrices for Human Evaluations

Figures 36, 37, 38 display the confusion matrices for the overall human evaluations on the Illusion part of each dataset. These matrices show the distribution of predictions made by human annotators for each true label.

14. Prompt Templates for GPT-4o

We use the following instructions to prompt GPT-4o.

For raw samples of IllusionMNIST, IllusionFashionMNIST, and IllusionAnimals datasets, we use the following prompt template:

> *"Which class is in the picture: {raw_class_names_str}. Just choose the correct class without any extra explanation."*

The raw class names are from the following labels:

- **MNIST**: digit 0, digit 1, digit 2, digit 3, digit 4, digit 5, digit 6, digit 7, digit 8, digit 9
- **Fashion-MNIST**: t-shirt/top, trouser, pullover, dress, coat, sandal, shirt, sneaker, bag, ankle boot

- (II): image on the left
- (RC): A bustling city street with neon lights and bustling crowd
- (IC / Ground Truth): elephant
- (Q): “There might be an illusion of something in the image or not. These are the classes that an illusion might belong to: {illusion_class_names_str}. Just choose the correct class without any extra explanation.”
- (A): elephant

Figure 4. An example of task definition

- (II): image on the left
- (RC): A bustling train station with passengers rushing to catch their trains
- (IC / Ground Truth): butterfly
- (Q): “There might be an illusion of something in the image or not. These are the classes that an illusion might belong to: {illusion_class_names_str}. Just choose the correct class without any extra explanation.”
- (A): butterfly

Figure 5. An example of task definition

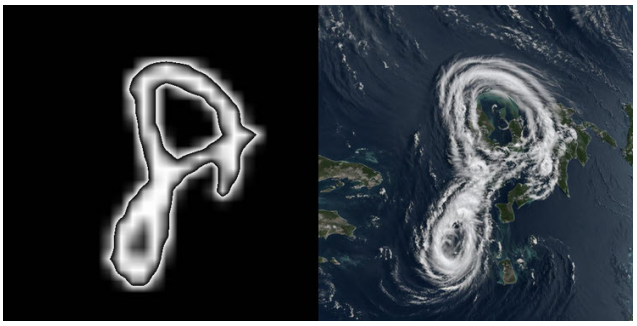

Figure 6. An example of IllusionMNIST

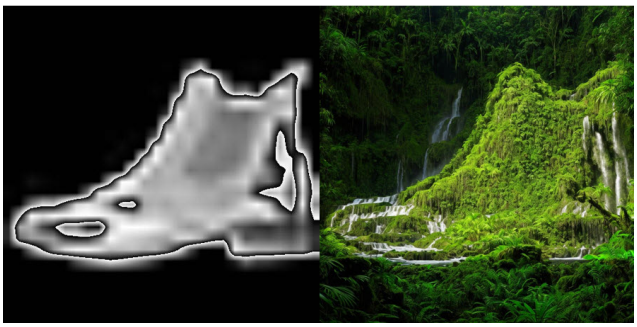

Figure 7. An example of IllusionFashionMNIST

Table 5. Human performance on 3 datasets: IllusionMNIST, IllusionFashionMNIST and IllusionAnimals

		Annotator 1	Annotator 2	Annotator 3	Annotator 4
Age		24	23	22	22
Gender		M	F	M	M
IllusionMNIST	**Accuracy**	98.35	95.04	-	96.69
	Precision	98.74	95.86	-	97.39
	Recall	97.73	94.41	-	96.02
	F1	98.05	94.77	-	96.47
IllusionFashionMNIST	**Accuracy**	73.02	69.84	80.95	-
	Precision	71.65	66.69	80.02	-
	Recall	71.55	69.27	80.07	-
	F1	70.95	67.33	79.34	-
IllusionAnimals	**Accuracy**	93.64	-	94.55	90.91
	Precision	92.95	-	95.38	92.98
	Recall	91.99	-	91.70	88.96
	F1	92.18	-	92.14	89.99

Table 6. Human performance on IllusionChar dataset

	Age	Gender	IllusionChar	
			WER	CER
Annotator 1	24	M	-	-
Annotator 2	23	F	32.22	14.11
Annotator 3	22	M	32.50	12.07
Annotator 4	22	M	31.11	13.78

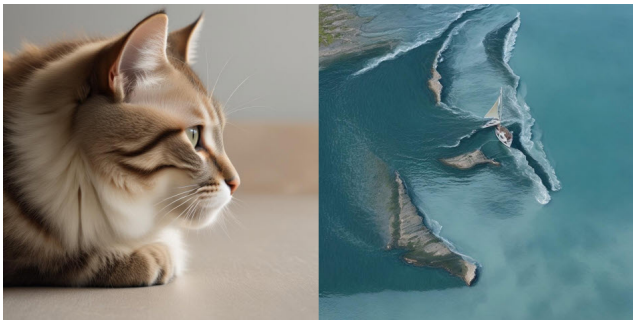

Figure 8. An example of IllusionAnimals

Figure 9. An example of IllusionChar

- **Animals**: cat, dog, pigeon, butterfly, elephant, horse, deer, snake, fish, rooster

For raw samples of the IllusionChar dataset, we use the following template:

> *"What sequence of characters are in the picture? Just say the sequence. Put your answer in quotation marks."*

For illusion and filtered samples of IllusionMNIST, IllusionFashionMNIST, and IllusionAnimals datasets, we use the following template:

> *"There might be an illusion of something in the image or not. These are the classes that an illusion might belong to: {illusion_class_names_str}. Just choose the correct class without any extra explanation."*

The illusion and filtered class names are the same as the raw class names with an additional "No illusion" class.

For illusion and filtered samples of the IllusionChar

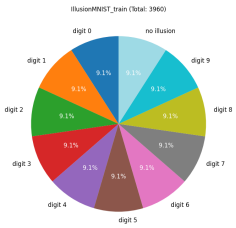

Figure 10. IllusionMNIST train split

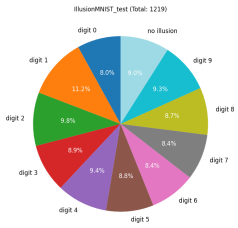

Figure 11. IllusionMNIST test split

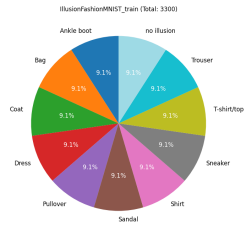

Figure 12. IllusionFashionMNIST train split

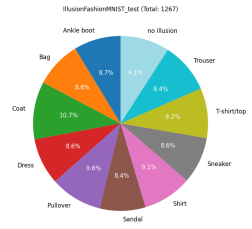

Figure 13. IllusionFashionMNIST test split

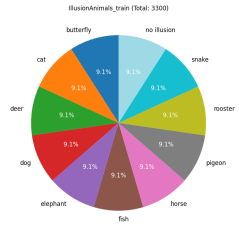

Figure 14. IllusionAnimals train split

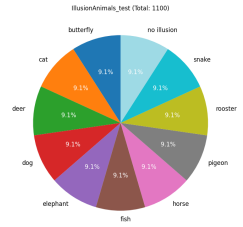

Figure 15. IllusionAnimals test split

dataset, we use the following template:

> *"There might be an illusion of a sequence of characters in the picture. If you cannot detect the sequence of characters, answer with "No illusion". If you can detect the sequence of characters, what sequence of characters are in the picture? Just say the sequence. Put your answer in quotation marks."*

15. GPT-4o Failed Examples

In this section, we present examples where GPT-4o correctly answered the Illusion image but failed to answer the Filtered image correctly. These examples are taken from four datasets: IllusionAnimals, IllusionMNIST, IllusionFashionMNIST, and IllusionChar. Examples are represented in the Figure 39.

16. Hyperparameters

The detailed hyperparameters for each model and dataset combination are summarized in Tables 7, 8.

17. Experimental Setup

For fine-tuning and inference of the CLIP, BLIP, and BLIP2 models, as well as for inference of the MiniGPT-V2 model on three datasets: IllusionMNIST, IllusionFashionMNIST, and IllusionAnimals, we use the Google Colab T4 GPU. For fine-tuning and inference of the LLaVA model, we use two T4 GPUs on Kaggle. For the Gemini and GPT-4o models, we use the official corresponding APIs.

18. Filter Details

Below, you can find the details of the filter we used to apply to illusory images for detecting illusions. We developed this filter, drawing inspiration from the human eye's ability to detect visual illusions when partially closed.

```python
import cv2
from google.colab.patches import cv2_imshow
import numpy as np

def generate_filtered_image(image_path, blur_amount=61):
    image = cv2.imread(f"{image_path}")
    blurred_image = cv2.GaussianBlur(image, (blur_amount, blur_amount), 0)
    blurred_image = cv2.blur(blurred_image, (20, 20))
    blurred_image = cv2.medianBlur(blurred_image, 5)
    gray_image = cv2.cvtColor(blurred_image, cv2.COLOR_BGR2GRAY)
    kernel = np.array([[-1, -2, -1], [-2, 13, -2], [-1, -2, -1]])
    sharpened_image = cv2.filter2D(gray_image, -1, kernel)
    return sharpened_image
```

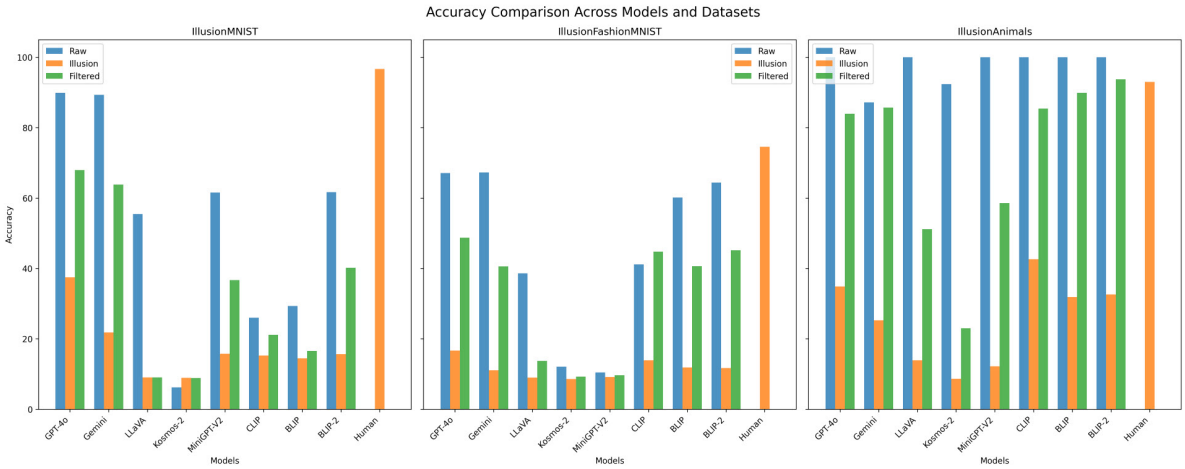

Figure 16. Visualization of zero-shot classification accuracies across various datasets

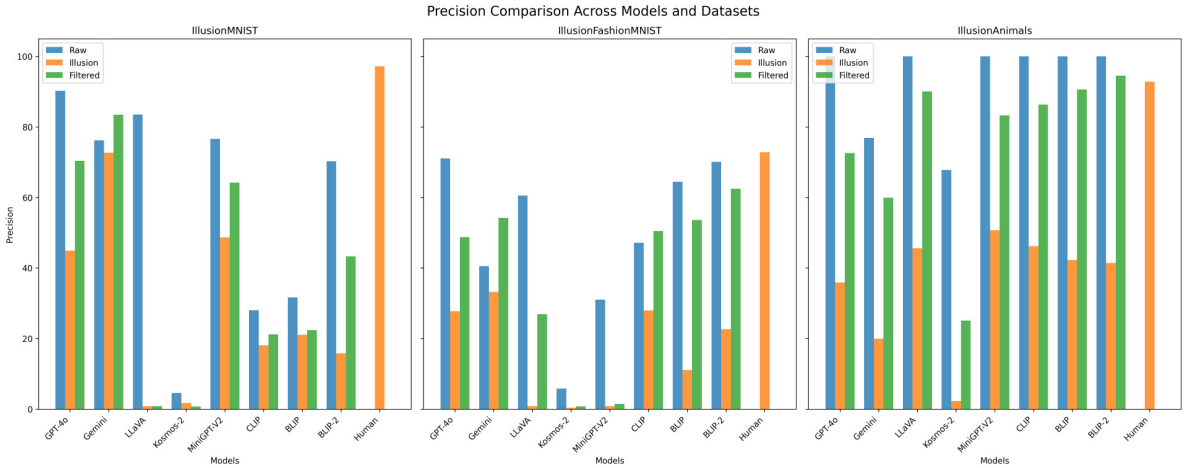

Figure 17. Visualization of zero-shot classification precisions across various datasets

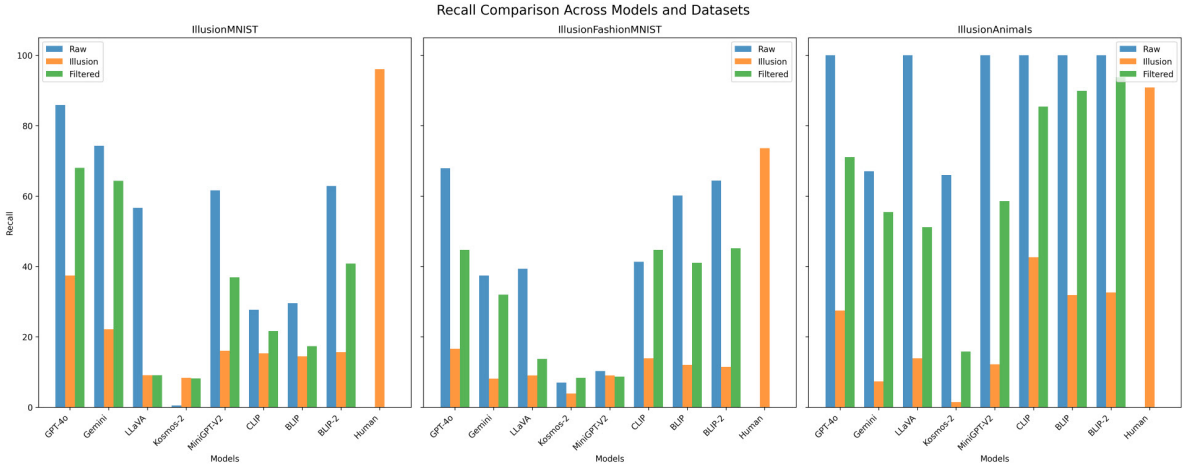

Figure 18. Visualization of zero-shot classification recalls across various datasets

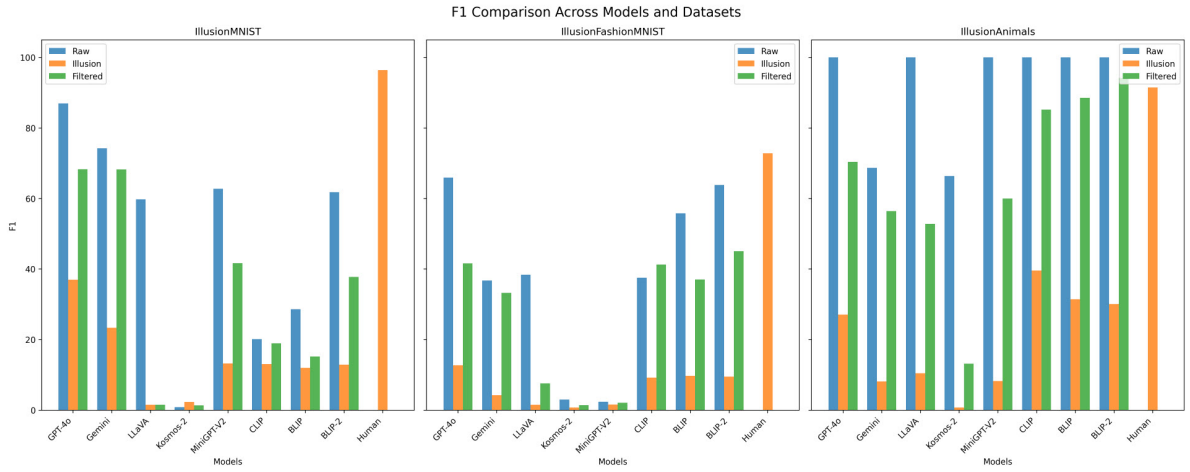

Figure 19. Visualization of zero-shot classification f1 scores across various datasets

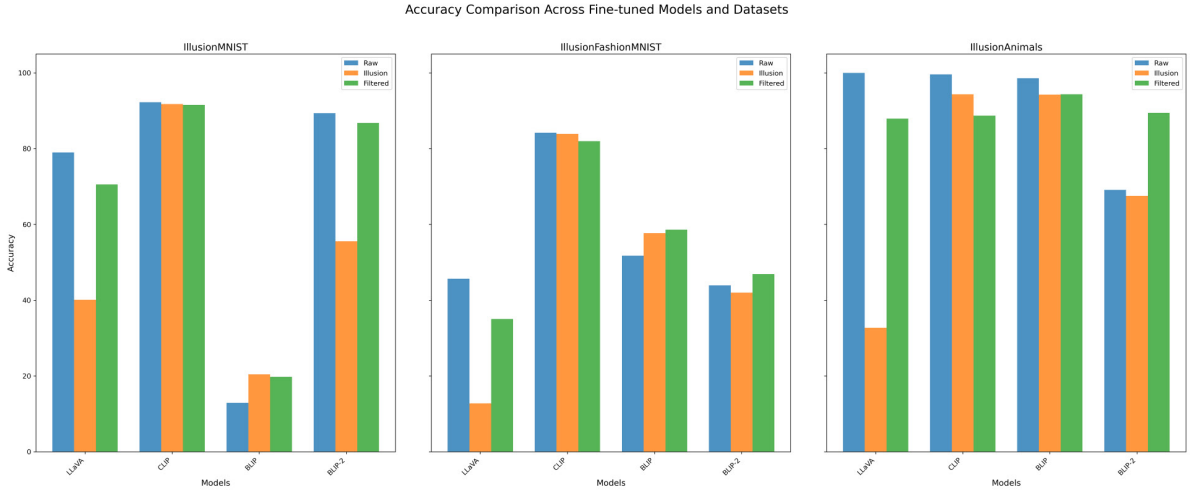

Figure 20. Visualization of fine-tuned classification accuracies across various datasets

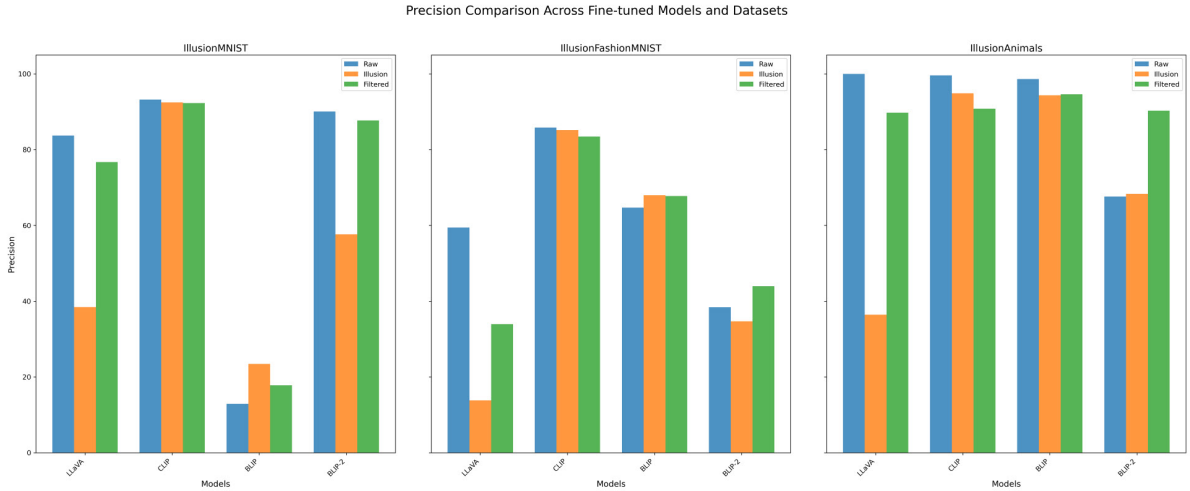

Figure 21. Visualization of fine-tuned classification precisions across various datasets

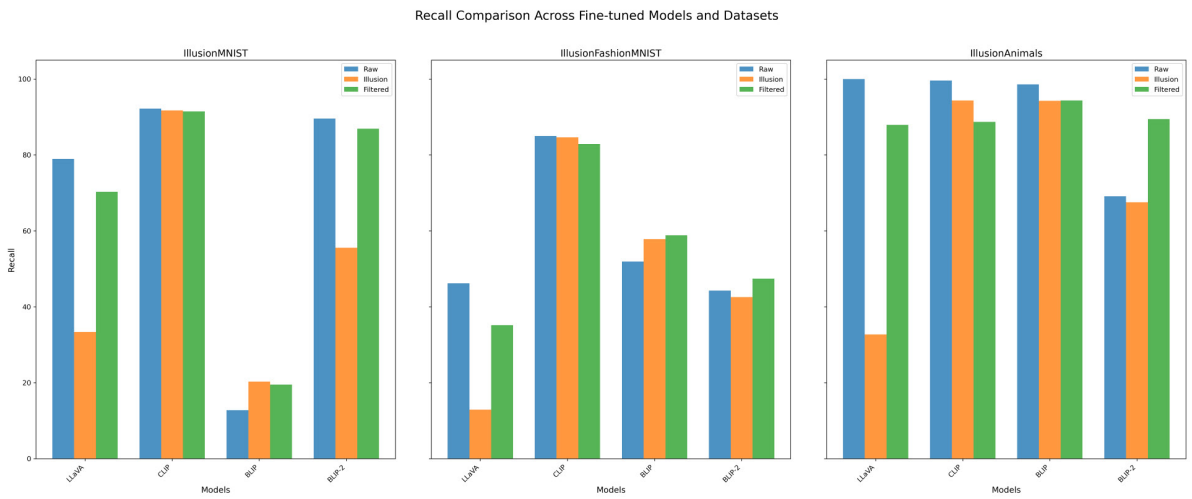

Figure 22. Visualization of fine-tuned classification recalls across various datasets

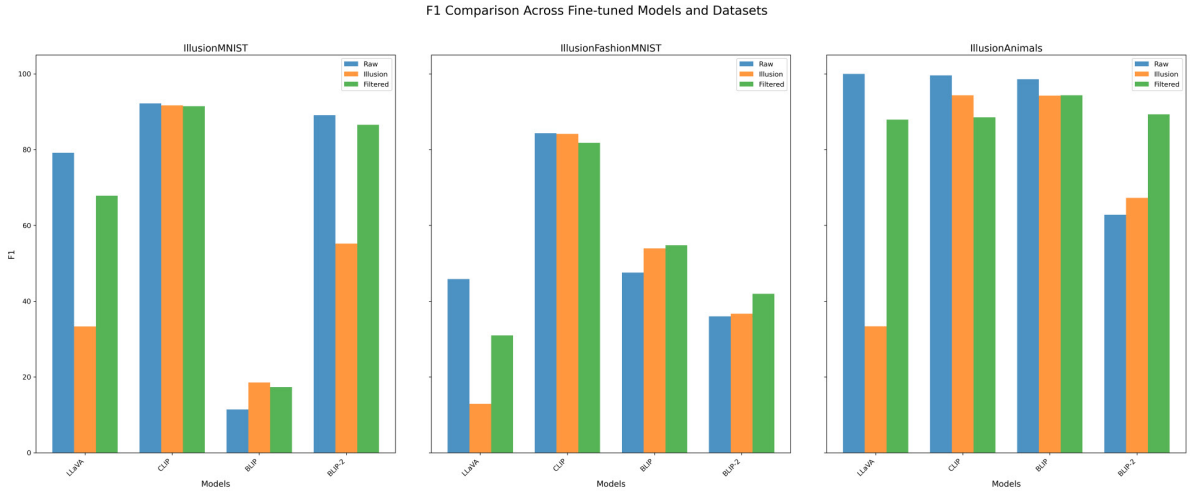

Figure 23. Visualization of fine-tuned classification f1 scores across various datasets

Table 7. Hyperparameters used for fine-tuning BLIP, BLIP-2, and CLIP Models on three datasets. All these experiments were conducted without a scheduler, using a train/test split of 90% and 10%, respectively, and a global seed of 10 to ensure reproducibility.

	Dataset	BLIP	BLIP-2	CLIP
Batch Size	**IllusionMNIST**	6	11	6
	IllusionFashionMNIST	6	11	6
	IllusionAnimals	6	11	6
Learning Rate	**IllusionMNIST**	1e-4	1e-5	1e-5
	IllusionFashionMNIST	5e-5	7e-5	1e-5
	IllusionAnimals	5e-5	1e-5	1e-5
Weight Decay	**IllusionMNIST**	1e-4	1e-5	1e-5
	IllusionFashionMNIST	1e-4	1e-5	1e-5
	IllusionAnimals	1e-4	1e-5	1e-5
# of Epochs	**IllusionMNIST**	3	3	3
	IllusionFashionMNIST	3	3	3
	IllusionAnimals	3	3	3
Optimizer	**IllusionMNIST**	AdamW	AdamW	AdamW
	IllusionFashionMNIST	AdamW	AdamW	AdamW
	IllusionAnimals	AdamW	AdamW	AdamW

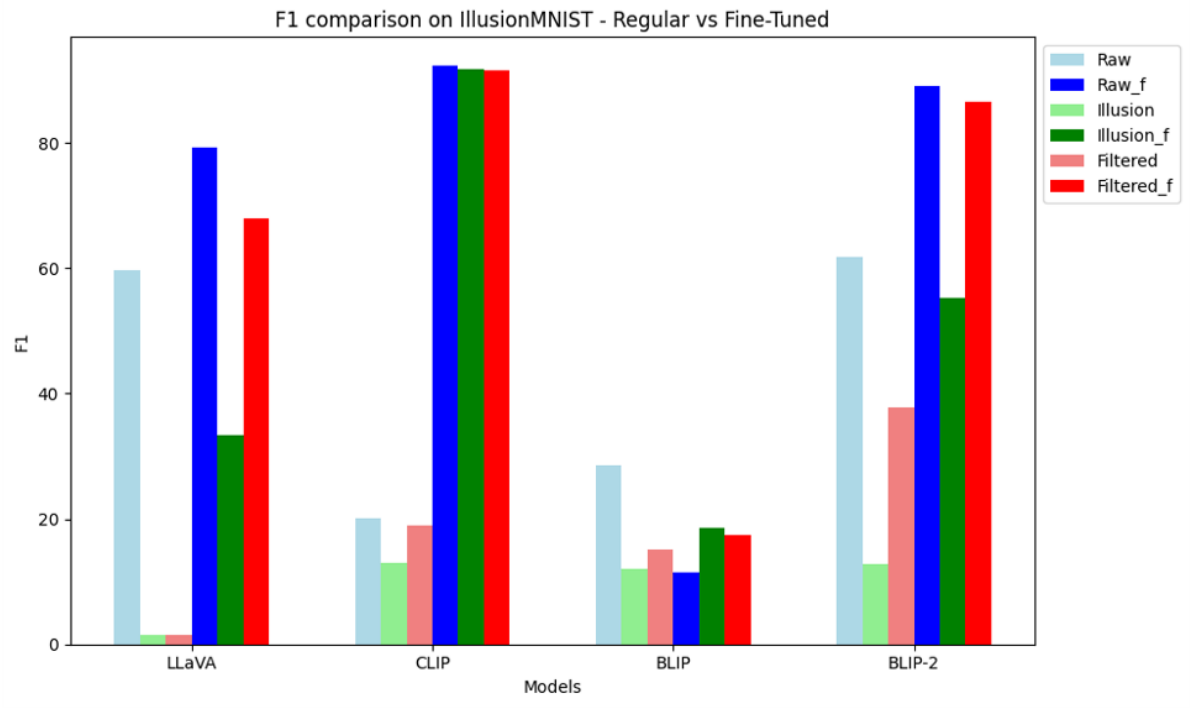

Figure 24. Comparison of zero-shot vs. fine-tuned classification f1 scores on IllusionMNIST dataset

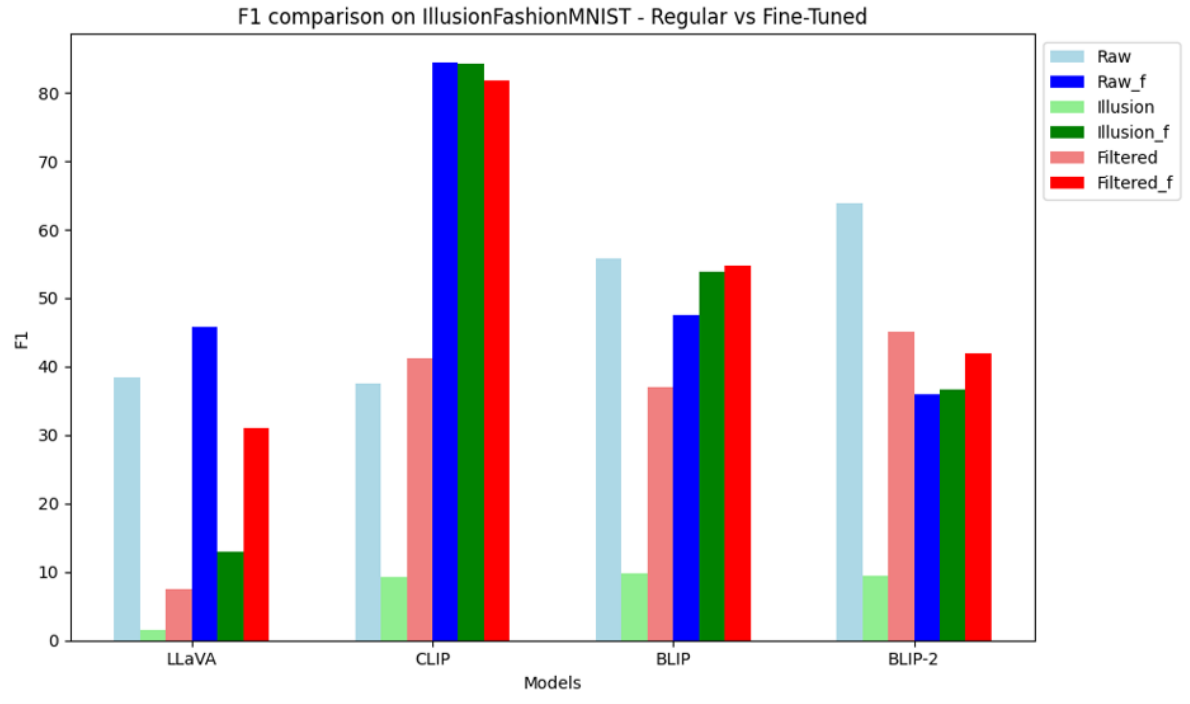

Figure 25. Comparison of zero-shot vs. fine-tuned classification f1 scores on IllusionFashionMNIST dataset

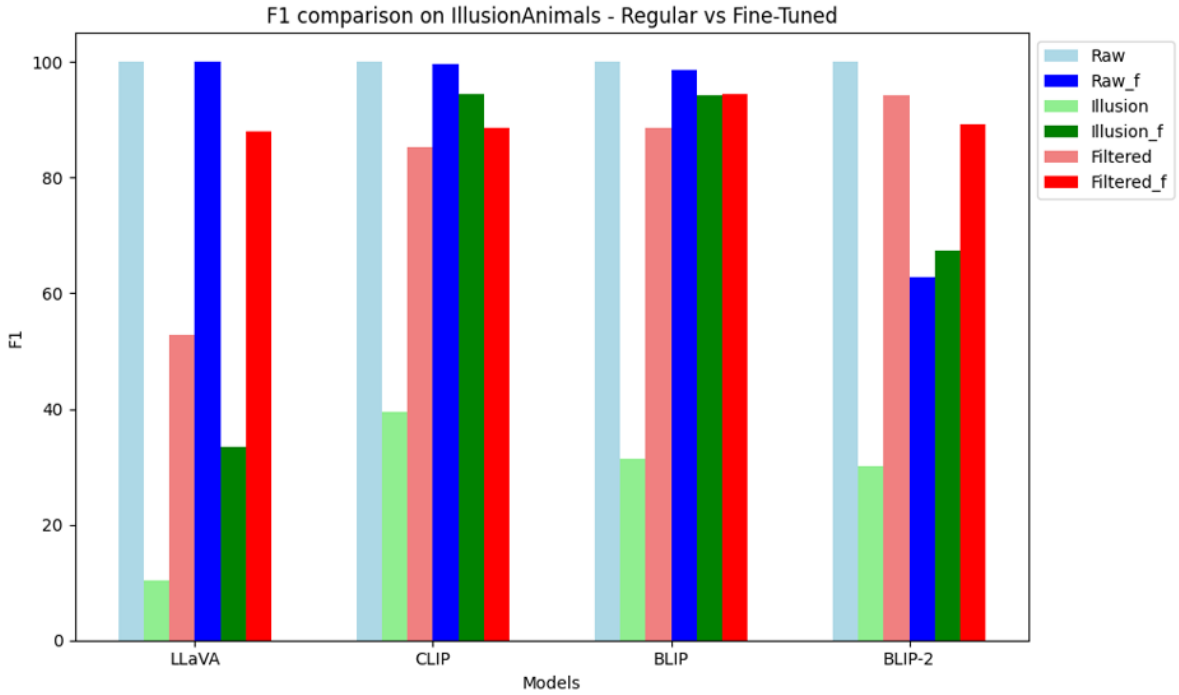

Figure 26. Comparison of zero-shot vs. fine-tuned classification f1 scores on IllusionAnimals dataset

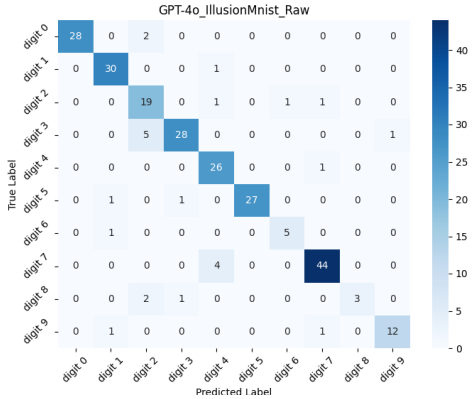

Figure 27. Confusion Matrix for GPT-4o on IllusionMNIST (Raw)

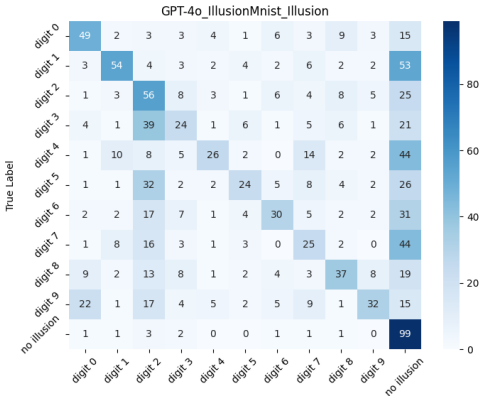

Figure 28. Confusion Matrix for GPT-4o on IllusionMNIST (Illusion)

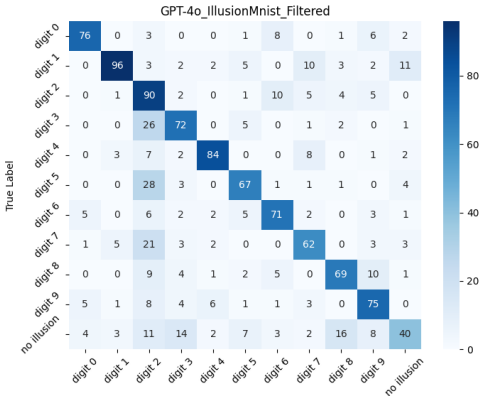

Figure 29. Confusion Matrix for GPT-4o on IllusionMNIST (Filtered)

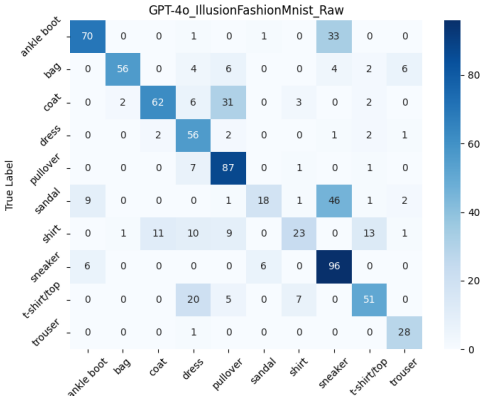

Figure 30. Confusion Matrix for GPT-4o on IllusionFashionMNIST (Raw)

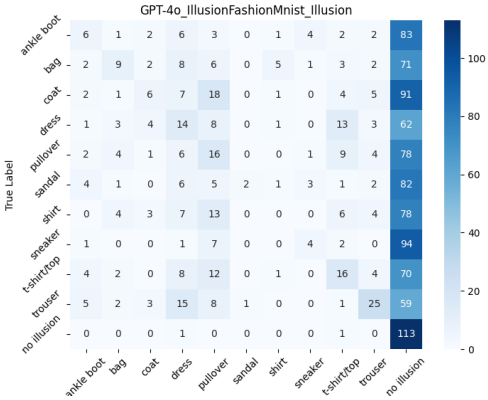

Figure 31. Confusion Matrix for GPT-4o on IllusionFashionMNIST (Illusion)

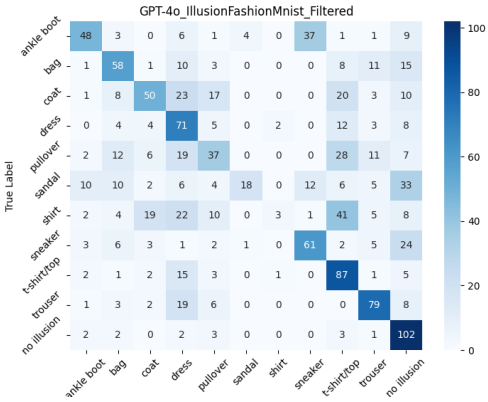

Figure 32. Confusion Matrix for GPT-4o on IllusionFashionMNIST (Filtered)

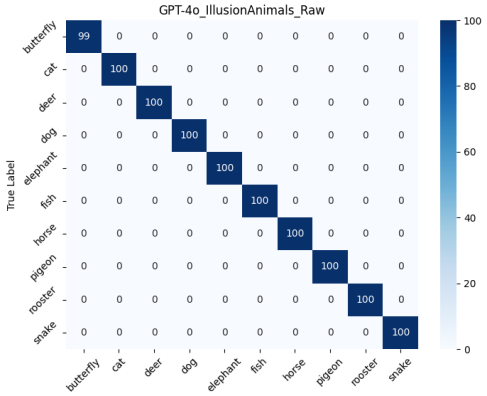

Figure 33. Confusion Matrix for GPT-4o on IllusionAnimals (Raw)

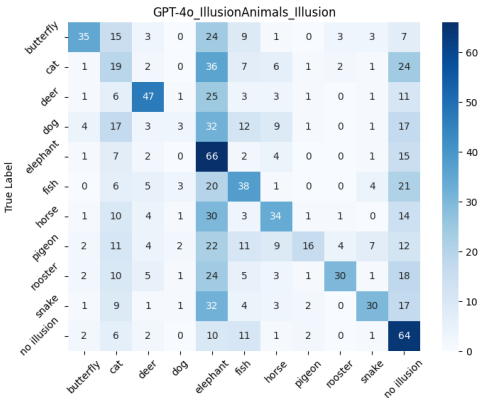

Figure 34. Confusion Matrix for GPT-4o on IllusionAnimals (Illusion)

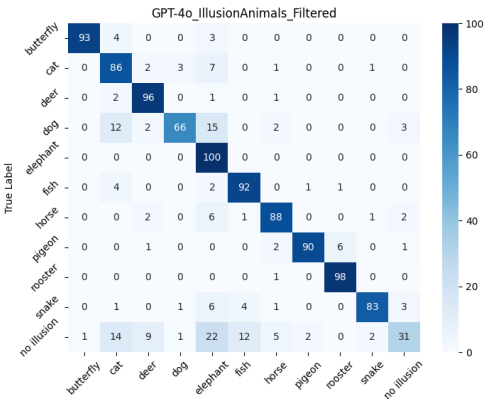

Figure 35. Confusion Matrix for GPT-4o on IllusionAnimals (Filtered)

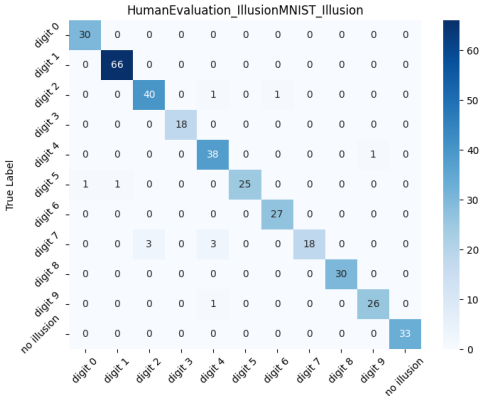

Figure 36. Confusion Matrix for Human Evaluation on IllusionMNIST (Illusion)

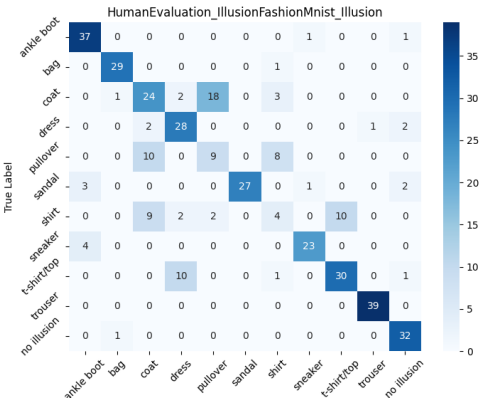

Figure 37. Confusion Matrix for Human Evaluation on IllusionFashionMNIST (Illusion)

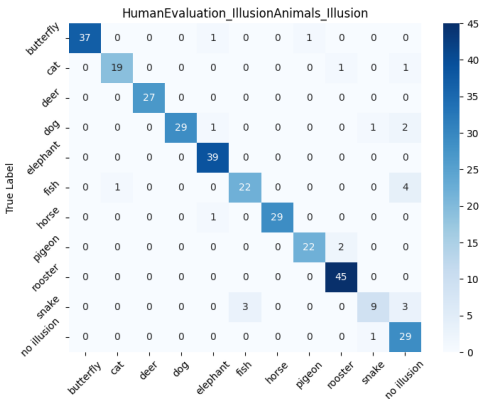

Figure 38. Confusion Matrix for Human Evaluation on IllusionAnimals (Illusion)

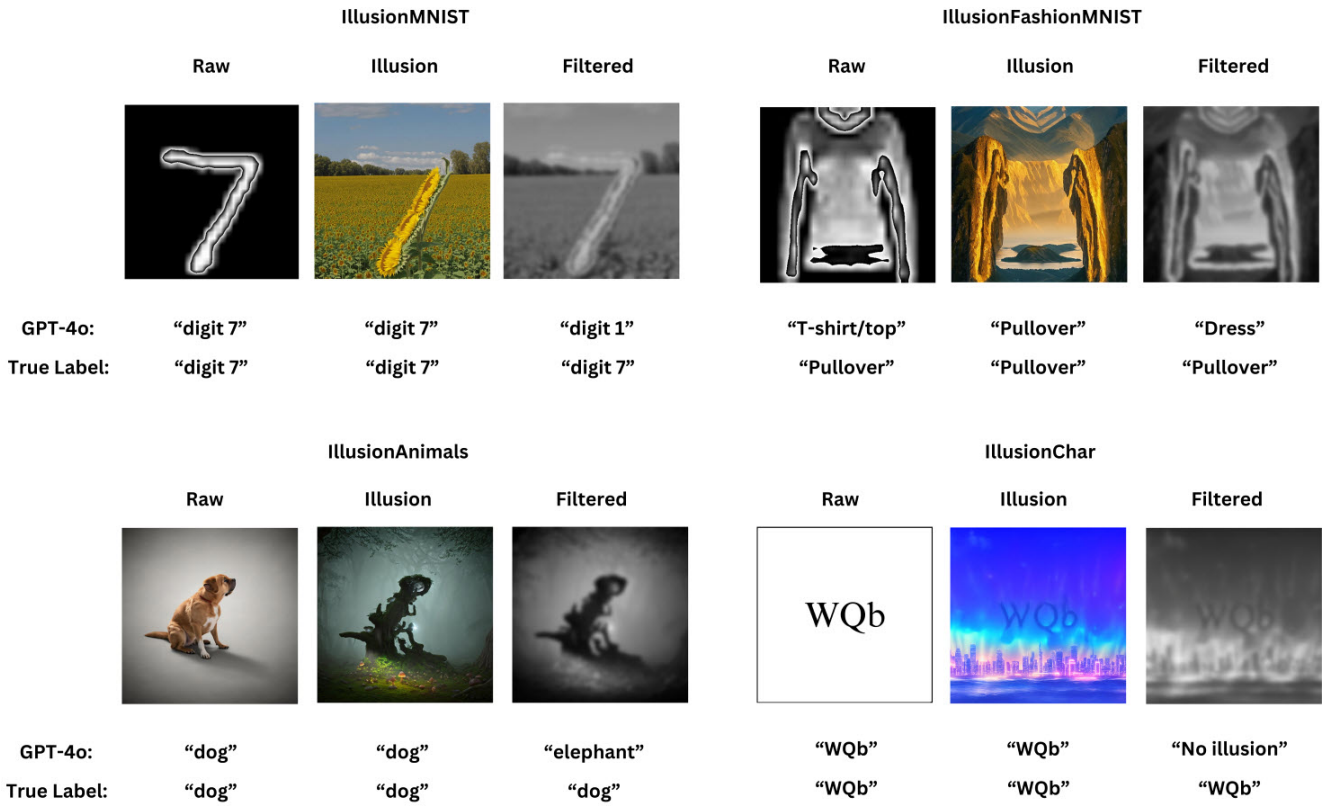

Figure 39. Examples of GPT-4o failing to correctly answer filtered images while correctly answering illusory images.

Table 8. Hyperparameters used for fine-tuning LLaVA model on three datasets, namely IllusionMNIST, IllusionFashionMNIST, and IllusionAnimals

Hyperparameter		**Value**
Learning Rate		1e-5
Batch Size (per device)		8
# of Epochs		2
Optimizer		AdamW
LoRA	**r**	64
	lora_alpha	16
	target_modules	all-linear
	dropout	0