# Diagnosing Vision Language Models' Perception by Leveraging Human Methods for Color Vision Deficiencies

**Kazuki Hayashi** **Shintaro Ozaki**
**Yusuke Sakai** **Hidetaka Kamigaito** **Taro Watanabe**
Nara Institute of Science and Technology (NAIST), Japan
hayashi.kazuki.hl4@is.naist.jp
ozaki.shintaro.ou6@naist.ac.jp
{sakai.yusuke.sr9, kamigaito.h, taro.watanabe}@is.naist.jp

## Abstract

Large-scale Vision-Language Models (LVLMs) are being deployed in real-world settings that require visual inference. As capabilities improve, applications in navigation, education, and accessibility are becoming practical. These settings require accommodation of perceptual variation rather than assuming a uniform visual experience. Color perception illustrates this requirement: it is central to visual understanding yet varies across individuals due to Color Vision Deficiencies, an aspect largely ignored in multimodal AI. In this work, we examine whether LVLMs can account for variation in color perception using the Ishihara Test. We evaluate model behavior through generation, confidence, and internal representation, using Ishihara plates as controlled stimuli that expose perceptual differences. Although models possess factual knowledge about color vision deficiencies and can describe the test, they fail to reproduce the perceptual outcomes experienced by affected individuals and instead default to normative color perception. These results indicate that current systems lack mechanisms for representing alternative perceptual experiences, raising concerns for accessibility and inclusive deployment in multimodal settings.

## 1 Introduction

Large-scale Vision-Language Models (LVLMs) (Liu et al., 2024; Abdin et al., 2024; Ye et al., 2024; Bai et al., 2025) extend language models with visual perception and enable unified multimodal reasoning. As LVLMs move from research prototypes to real-world interactive systems, they are increasingly deployed in navigation, education, and assistive technologies (Zitkovich et al., 2023; Hao et al., 2024; Morita et al., 2024; Zhou et al., 2025). In such applications, LVLMs serve as a medium through which users access and interpret visual information in their environment.

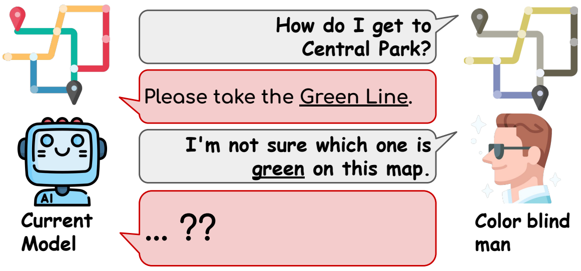

How do I get to Central Park?
Please take the Green Line.
I'm not sure which one is green on this map.
... ??
AI
Current Model
Color blind man

Figure 1: An example of communication failure due to CVDs: a user with CVD may not correctly interpret the model's color-based instructions. This highlights the need for LVLMs to account for perceptual differences in real-world interactions.

A key yet underexplored assumption in such deployments is that users share a similar visual experience of the world. In reality, visual perception varies across individuals. Color vision in particular shows substantial variability due to biological factors (Emery and Webster, 2019), and Color Vision Deficiencies (CVDs) are a common form of such diversity, impairing discrimination between colors such as red and green (Birch, 2012; Roberson and Hanley, 2007). These differences are not only perceptual but also functional: color vision affects wayfinding, diagram and chart reading, medical interpretation, and user interface navigation. As illustrated in Figure 1, an LVLM-based navigation assistant instructing a traveler to "Take the green line" may fail to provide actionable information for a red–green color blind user, raising concerns about inclusion, accessibility, and fairness in real deployments (Kawakita et al., 2024). Systems designed around a "standard" visual experience may silently exclude users with different perceptual characteristics, even when linguistic content remains fully accessible.

Current LVLMs are trained largely on web-scale image-text corpora that implicitly assume typical color vision (Paik et al., 2021; Samin et al., 2025; Rahmanzadehgervi et al., 2024). As a result, these

*Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics*
*Volume 1: Long Papers*, pages 7582–7605
March 24-29, 2026 ©2026 Association for Computational Linguistics

models acquire substantial knowledge about color, color blindness, and clinical diagnostic tools. However, whether such knowledge enables the model to simulate alternative visual percepts remains unclear. This distinction is central: explaining what CVD is in language is not the same as perceiving colors as a CVD user would. To our knowledge, the ability of LVLMs to reproduce perceptual diversity has not been systematically investigated.

To address this, we leverage the Ishihara Test (Ishihara, 1917), a clinically validated tool widely used for diagnosing CVDs in humans. Rather than using Ishihara solely as a screening instrument, we repurpose it as a *diagnostic probe* for LVLMs. Because Ishihara plates are designed to elicit different perceptual responses across color vision types, they allow direct comparison between model outputs and established human perceptual categories. In addition, the test provides a standardized visual stimulus that isolates color-based perceptual variation without cultural or linguistic interference, and its clinical validation over a century allows perceptual differences to be evaluated without new annotation pipelines or subjective rating procedures.

We evaluate LVLM behavior through three complementary perspectives. At the *generation level*, we assess whether the model outputs the digit perceived by users with different CVD types. At the *confidence level*, we analyze perplexity under simulated CVD conditions to determine whether uncertainty patterns align with human perceptual tendencies. At the *representation level*, we apply a layer-wise LogitLens probing (nostalgebraist, 2020) method to assess whether internal activations encode distinctions between vision types. Together, these perspectives allow us to analyze how perceptual variation interacts with LVLMs' generative, probabilistic, and latent processing mechanisms.

Our findings show that although LVLMs possess strong linguistic knowledge about CVDs and can describe the Ishihara Test, they fail to reproduce the perceptual outcomes experienced by individuals with CVD. Even when given additional linguistic descriptions or visual exemplars, the models default to normative color perception, indicating insufficient alignment between linguistic knowledge and perceptual grounding. These results suggest that current LVLMs lack mechanisms for representing alternative perceptual experiences, which has implications for accessibility, fairness, and human-centered deployment in multimodal AI as such systems increasingly mediate visual information.

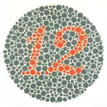
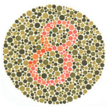
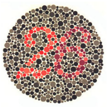
| Plate | Type | Answer |
|---|---|---|
|  | Normal | 12 |
| | Protanopia | 12 |
| | Deuteranopia | 12 |
| | Tritanopia | 12 |
|  | Normal | 8 |
| | Protanopia | 3 |
| | Deuteranopia | 3 |
| | Tritanopia | 8 |
|  | Normal | 26 |
| | Protanopia | 6 |
| | Deuteranopia | 2 |
| | Tritanopia | 26 |

Table 1: Ground-truth answers for each Ishihara plate under different vision conditions, based on clinical interpretations of the Ishihara color test (Ishihara, 1917).

## 2 Background

### 2.1 Color Vision Deficiency and Ishihara Test

Color vision in humans relies on three types of cone photoreceptors: red, green, and blue, each sensitive to long, medium, or short wavelengths. CVDs affect about 8% of males and 0.5% of females (Birch, 2012). When one type of cone is missing or doesn't work properly, large parts of the color spectrum lose their typical hue or brightness, resulting in CVDs. The three main types are summarized below:

**Protanopia** Absence of long wavelength (L) cones makes reds appear very dark and compresses the differences among many reds, greens, and browns; purples are often mistaken for blue.

**Deuteranopia** Absence of medium wavelength (M) cones leaves brightness almost unaffected but shifts greens toward beige, producing strong red and green confusion.

**Tritanopia** Absence of short wavelength (S) cones narrows the blue to yellow axis; blues drift toward green, cyans toward gray, and yellows toward pink or light gray, greatly reducing blue versus yellow separability.

To screen for CVDs, the Ishihara Test (Ishihara, 1917) has been widely used for decades in clinical and research settings (Fanlo Zarazaga et al., 2019). Each plate shows a figure made of colored dots. People with normal vision usually identify it, while those with CVDs may see a different number or nothing. For instance, the plate in Table 1 looks like "8" to a person with normal vision, but is often read

| Input | Extracted mPLUG-Owl3 Output |
|---|---|
| What is Ishihara Test? | ⋯ visual test **used to detect color-vision deficiencies, especially red–green types such as protanopia and deuteranopia.** It presents a pattern of red and green dots forming a hidden number; individuals with normal vision can read the number, whereas those with certain CVDs cannot. |
| What is Protanopia? | ⋯ **color blindness that affects a person's ability to distinguish between red and green colors.** In people with protanopia, the cone cells in the retina that are responsible for perceiving red and green colors are not functioning properly. |

Table 2: Excerpts from mPLUG-Owl3 showing knowledge of the Ishihara test and CVDs. The full outputs are available in Appendix B.

as "3" by people with red and green deficiencies (Birch, 2012). By using typical color confusions across CVDs, the Ishihara Test remains a fast and reliable tool for identifying these conditions.

### 2.2 CVDs and Ishihara Tests in Recent Research

Despite its limited size, the Ishihara Test remains a clinically validated and reliable tool for CVD detection. Chen and Hsu (1995) applied automated analysis to recognize Ishihara patterns, enabling automated color vision testing. Later work confirmed that digital versions retain reliability across devices (Zhao et al., 2019; Klinke et al., 2024). Henshaw et al. (2025); Grover et al. (2025) introduced synthetic Ishihara-style datasets to assess model performance on digit or character recognition under red and green confusion. CVDs compensation methods also use Ishihara principles to enhance accessibility for people with color vision deficiency. Key approaches include physiologically based daltonization (Machado et al., 2009), eigenvector-based recoloring (Lin et al., 2019), and neural real-time Ishihara recoloring (Montes Rivera et al., 2020). These works highlight Ishihara plates as diagnostic tools and foundations for assistive color algorithms.

### 2.3 Color Understanding in LVLMs

Early VQA models (Goyal et al., 2017) explored color-related questions but relied on priors over visual understanding, yielding inconsistent results. Agrawal et al. (2018); Kervadec et al. (2021) proposed debiased datasets to reduce language bias and improve visual grounding. Transformer-based models (Radford et al., 2021; Kim et al., 2021) enhanced image-text alignment and performed well on some tasks yet still struggle with fine-grained color distinctions (Samin et al., 2025). Although recent LVLMs exhibit strong visual reasoning, their color understanding remains limited by insufficient supervision for subtle or ambiguous cases. Burapacheep et al. (2024); Shahgir et al. (2024) found that models struggle with basic color-word grounding and are easily misled by perceptual illusions. Liang et al. (2025) evaluated diverse color tasks and showed even large models struggle with shifts and ambiguity, revealing persistent limitations.

## 3 Experimental Setup

### 3.1 Pilot Study

Before analyzing our main question of whether LVLMs can simulate aspects of human color perception, we first verified that they possess basic language-level knowledge about color vision deficiencies and the Ishihara Test. To test this prerequisite, we used two types of prompts: (i) asking an Ishihara plate "`What is this image used for?`", and (ii) requesting definitions such as "`What is Protanopia as a type of color vision deficiency?`". As shown in Table 2, all evaluated models were judged to respond correctly. This confirms that they satisfy the necessary prerequisite knowledge for the analyses that follow. Appendix B provides full outputs.

### 3.2 Research Questions

We use the Ishihara Test as with humans to analyze how LVLMs simulate color vision across task settings[1]. In our research, "color perception" means the model's ability to respond like humans with specific CVDs, rather than identifying colors. We investigate this across three levels: generation, confidence, and internal representation.

**RQ1 (Generation level).** Can LVLMs produce Ishihara digit responses that align with those reported by individuals with color vision deficiency types? This evaluates whether the model' s generative outputs can emulate alternative perceptual outcomes rather than defaulting to neurotypical perception.

**RQ2 (Confidence level).** Do model confidence patterns vary across vision conditions, and do

[1] We use 25 out of 38 plates that require numeral responses to ensure consistent analysis across visual conditions.

| Type | Prompt |
|---|---|
| **Base** | {Image: Ishihara plate}<br>You are Protanopic.<br>What number do you see? Only provide the answer. No additional information. |
| **Linguistic Support** | {Image: Ishihara plate}<br>You are Protanopic.<br>Protanopia means you have difficulty distinguishing red tones, as red appears dimmer and may be confused with green or brown.<br>What number do you see? Only provide the answer. No additional information. |
| **Visual Support** | {Reference Image: Ishihara plate}<br>You are Protanopic.<br>Example 1: Image shows an Ishihara plate. A Protanopic person sees this number: 2.<br>{Image: Ishihara plate}<br>Now, look at the following image. What number do you see? Only provide the answer. No additional information. |
| **Doctor-Style** | {Image: Ishihara plate}<br>When a person sees “{task_number}” in this Ishihara test plate, which type of color vision is most likely?<br>Choose from the following options:<br>- Normal - Protanopia - Deuteranopia - Tritanopia Please output your answer. |

Table 3: Examples of prompts used in our experiments. Base, Linguistic, and Visual prompts are used in RQ1-RQ2 for digit generation, and Doctor-Style is used in RQ3 for vision-type inference. All include an Ishihara plate as visual input.

uncertainty signals (measured via perplexity) reflect condition-specific perceptual ambiguity observed in humans? This probes whether LVLMs assign different likelihoods to Ishihara responses depending on simulated vision type.

**RQ3 (Representation level).** Do internal layer-wise activation patterns encode distinctions consistent with color vision deficiency categories? This examines whether alternative perceptual states are represented at the latent level, even if not expressed in the final output.

### 3.3 Task Definition

The task is inspired by the human Ishihara Test: given an Ishihara plate, the model is instructed to simulate a target vision condition and report the perceived digit. We simulate four conditions: {Normal, Protanopia, Deuteranopia, Tritanopia}. Since the Ishihara Test does not assess Tritanopia, the correct digits for Normal and Tritanopia coincide. We exploit this property by treating Tritanopia as a *control condition*: if a model meaningfully incorporates the instruction “You are Tritanopic,” its output distribution or confidence should differ from Normal despite identical ground-truth digits. As shown later, most LVLMs instead return nearly identical responses for the two conditions, suggesting weak sensitivity to instructed perceptual states.

**Prompts** To analyze whether LVLMs can simulate CVDs under different instructional prompts, we use Ishihara plates and design four types of prompts: Base, Linguistic Support, Visual Support and Doctor-Style, as shown in Table 3. We tested multiple variants within each prompt type but observed no substantial differences. Thus, we use one representative set in our main experiments and report the variant analysis in Appendix C.

**Base Prompts** These provide only the condition, e.g., “You are Protanopic”, and ask the model to report the number in the Ishihara plate. This tests whether the model can simulate CVD perception specified by the condition.

**Linguistic Support** In addition to the Base prompt, these add a short description of the impairment, e.g., “red tones appear darker,” to assess whether textual context aids prediction.

**Visual Support** These include a brief condition description and a reference example showing what number appears in another plate, before asking the model to identify the number in a new image. This tests whether few-shot visual examples help simulate CVD perception.

**Doctor-Style** Instead of generating digits, the model infers the vision type (Normal, Protanopia, Deuteranopia, or Tritanopia) from a given response. Used in RQ3, this prompt avoids tokenization issues with multi-digit numerals (e.g., “18” → “1”, “8”) by using diagnosis labels, which are consistently tokenized and clearly distinguishable across models.

### 3.4 Evaluation Metrics

We quantify each research question using a dedicated and interpretable metric for each aspect.

**Digit Accuracy (RQ1)** The proportion of Ishihara plates on which the model’s generated digit matches the ground-truth numeral under the specified vision condition.

| LVLM | Size | Normal | | | Protanopia | | | Deuteranopia | | | Tritanopia | | | Avg. | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **Base** | **Ling.** | **Vis.** | **Base** | **Ling.** | **Vis.** | **Base** | **Ling.** | **Vis.** | **Base** | **Ling.** | **Vis.** | **Base** | **Ling.** | **Vis.** |
| Llama-3.2 | 11B | 14.3 | 9.5 | – | 5.9 | 5.9 | – | 11.8 | 5.9 | – | 4.8 | 23.8 | – | 9.2 | 11.3 | – |
| LLaVA-NeXT | 13B | 52.4 | 52.4 | – | 5.9 | 5.9 | – | 5.9 | 5.9 | – | 52.4 | 61.9 | – | 29.2 | 31.5 | – |
| mPLUG-Owl3 | 7B | 71.4 | 71.4 | 62.0 | 5.9 | 5.9 | 5.9 | 5.9 | 5.9 | 17.7 | 66.7 | 66.7 | 62.0 | 37.5 | 37.5 | 36.9 |
| Phi-3.5 | 4.2B | 0.0 | 0.0 | 4.8 | 0.0 | 0.0 | 5.9 | 0.0 | 0.0 | 11.8 | 0.0 | 0.0 | 9.5 | 0.0 | 0.0 | 8.0 |
| Qwen2.5-VL | 7B | 28.6 | 0.0 | 4.8 | 0.0 | 0.0 | 5.9 | 0.0 | 0.0 | 0.0 | 23.8 | 0.0 | 4.8 | 13.1 | 0.0 | 3.9 |
| GPT-4o | – | 90.5 | 90.5 | 90.5 | 23.6 | 17.7 | 23.6 | 5.9 | 23.5 | 5.9 | 71.4 | 66.7 | 42.9 | 47.9 | 48.4 | 40.7 |

Table 4: *Digit Accuracy* (%) of each LVLM under simulated Vision types: **Normal** and three CVD types, with support settings: **Base**, **Ling.** (Linguistic), and **Vis.** (Visual). **Vis.** was used only for models supporting multi-image input. Plates for which no correct digit exists (marked as "N/A"in the Gold answers in Table 5) were excluded from the accuracy calculation.

**Per-Token Perplexity (RQ2)** For each plate, we concatenate the image, prompt, and gold numeral answer into a single input sequence. All tokens preceding the answer are masked with `ignore_index`, ensuring that only the answer tokens contribute to the loss. The answer consists of $L$ tokens, denoted as $\boldsymbol{x}_{1:L}$. We compute their token-level log-probabilities

$$v_i = \log P\left(x_i \mid \textit{prompt}, \boldsymbol{x}_{1:i-1}\right),$$

and the average negative log-likelihood

$$\mathcal{L} = -\frac{1}{L}\sum_{i=1}^{L} v_i, \quad \text{PPL} = \exp(\mathcal{L}).$$

A lower PPL indicates higher confidence in the forced-decoded answer.

**Layer-Wise Diagnosis Probability (RQ3)** This approach builds on prior work showing that intermediate representations influence model predictions (Wendler et al., 2024; Schut et al., 2025). At each Transformer layer, we project the hidden state $\boldsymbol{h}$ into the vocabulary space using the unembedding matrix $\mathbf{W}^{\text{unembed}}$ and apply a softmax:

$$\text{LogitLens}(\boldsymbol{h}) = \text{softmax}\left(\mathbf{W}^{\text{unembed}}\,\boldsymbol{h}\right).$$

The resulting distribution gives the probability of every token at that layer (nostalgebraist, 2020). For a specific diagnosis-label token $t$ (e.g., *Protanopia* or *Deuteranopia*) we record

$$P(t) = \frac{\exp\left(\boldsymbol{W}_t^{\text{unembed}} \cdot \boldsymbol{h}\right)}{\sum_j \exp\left(\boldsymbol{W}_j^{\text{unembed}} \cdot \boldsymbol{h}\right)}.$$

Plotting $P(t)$ across layers reveals how strongly the model favors each color-vision diagnosis while processing the plate, offering insights into its internal decision-making process.

### 3.5 Models

We evaluate six LVLMs: Phi-3.5 (Abdin et al., 2024), Qwen2.5-VL (Bai et al., 2025), mPLUG-Owl3 (Ye et al., 2024), LLaVA-NeXT (Liu et al., 2024), Llama-3.2 (Grattafiori et al., 2024), and GPT-4o (OpenAI et al., 2024). GPT-4o is proprietary; the remaining five are open-source and were selected for passing the Pilot Study prerequisite. Appendix A provides the detailed settings.

## 4 Results and Analysis

### 4.1 RQ1: Generation Level

We use *Digit Accuracy* to evaluate how often the model correctly identifies the numeral on each Ishihara plate. Table 4 shows accuracy across conditions and prompt types. In the Normal condition, humans are expected to score 100 %, but only GPT-4o approaches this with 90.5 %, followed by mPLUG-Owl3 at 71 %, while some models score near 0 %. Performance collapses for both Protanopia and Deuteranopia, with no model exceeding 24 %, clearly indicating the difficulty of simulating CVDs. With Linguistic Support, brief descriptions shift accuracy by at most two points without changing rankings; GPT-4o improves to 23 % on Deuteranopia but drops on Tritanopia. Qwen2.5-VL often refuses under disability-related terminology, averaging 0 %. With Visual Support, few-shot examples have minimal effect; accuracy shows no consistent improvement and sometimes decreases.

Table 5 lists the gold digits and Base-prompt predictions across 25 plates. GPT-4o reproduces most digits under Normal vision, and mPLUG-Owl3 performs well, but the remaining models are inconsistent. Under Protanopia and Deuteranopia, most models repeat Normal outputs, return empty

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
|  | **Gold** | **12** | **8** | **6** | **29** | **57** | **5** | **3** | **15** | **74** | **2** | **6** | **97** | **45** | **5** | **7** | **16** | **73** | **N/A** | **N/A** | **N/A** | **N/A** | **26** | **42** | **35** | **96** |
| Normal | Llama | 19 | **8** | 8 | 42 | 7 | 1 | 2 | 4 | 15 | **2** | 42 | 42 | 1 | 7 | **7** | 4 | 53 | 15 | 15 | 3 | 4 | 52 | 52 | 52 | 8 |
|  | LLaVA | **12** | **8** | **6** | **29** | 37 | 3 | 6 | 16 | 7 | **2** | **6** | 9 | 5 | 6 | 7 | **16** | 3 | 1 | 1 | 1 | 1 | 28 | **42** | **35** | **96** |
|  | mPLUG | **12** | **8** | **6** | **29** | 37 | **5** | **3** | **15** | 24 | **2** | **6** | 9 | 10 | **5** | **7** | **16** | 13 | 100 | 100 | 10 | 100 | 23 | **42** | **35** | **96** |
|  | Phi | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – |
|  | Qwen | **12** | **8** | **6** | **29** | **57** | 74 | 74 | 74 | **74** | 74 | 74 | 74 | 74 | 74 | 74 | 74 | 74 | 74 | 74 | 74 | 74 | 74 | 48 | 74 | 95 |
|  | GPT | **12** | **8** | **6** | **29** | **57** | **5** | **3** | **15** | **74** | **2** | **6** | 74 | **45** | **5** | **7** | **16** | **73** | 74 | 74 | 74 | 74 | 5 | **42** | **35** | **96** |
|  | **Gold** | **12** | **3** | **5** | **70** | **35** | **2** | **5** | **17** | **21** | **N/A** | **N/A** | **N/A** | **N/A** | **N/A** | **N/A** | **N/A** | **N/A** | **5** | **2** | **45** | **73** | **6** | **2** | **5** | **6** |
| Protanopia | Llama | 19 | 2 | **5** | 17 | 3 | 3 | 2 | 6 | 3 | 2 | 2 | 2 | 45 | 7 | 4 | 6 | 46 | 73 | 4 | 53 | 52 | 4 | 4 | 4 | 2 |
|  | LLaVA | **12** | 8 | 6 | 29 | 57 | 3 | 6 | 16 | 7 | 2 | 6 | 9 | 1 | 6 | 7 | 16 | 3 | 1 | 1 | 1 | 1 | 20 | 42 | 38 | 96 |
|  | mPLUG | **12** | 8 | 6 | 29 | 37 | 5 | 3 | 15 | 24 | 9 | 6 | 9 | 10 | 5 | 7 | 16 | 13 | 1 | 1 | 1 | 1 | 23 | 42 | 35 | 96 |
|  | Phi | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – |
|  | Qwen | 21 | 8 | 9 | 29 | 74 | 74 | 74 | 74 | 74 | 74 | 74 | 74 | 74 | 74 | 74 | 74 | 7 | 74 | 74 | 74 | 74 | 38 | 43 | 74 | 95 |
|  | GPT | **12** | **3** | 3 | 29 | 55 | 3 | **5** | 15 | 4 | 9 | 3 | 37 | 45 | **N/A** | 7 | 16 | 13 | 74 | **2** | 74 | 27 | 25 | **N/A** | 39 | 96 |
|  | **Gold** | **12** | **3** | **5** | **70** | **35** | **2** | **5** | **17** | **21** | **N/A** | **N/A** | **N/A** | **N/A** | **N/A** | **N/A** | **N/A** | **N/A** | **5** | **2** | **45** | **73** | **2** | **4** | **3** | **9** |
| Deuteranopia | Llama | 19 | 2 | **5** | 52 | 1 | 53 | 52 | 49 | 45 | 53 | 52 | 52 | 36 | 55 | 7 | 56 | 4 | 4 | **2** | 49 | 53 | 42 | 52 | 37 | 8 |
|  | LLaVA | **12** | 8 | 6 | 29 | 57 | 3 | 3 | 16 | 14 | 2 | 6 | 9 | 10 | 6 | 7 | 16 | 3 | 10 | 1 | 1 | 12 | 20 | 42 | 38 | 96 |
|  | mPLUG | **12** | 8 | 6 | 29 | 37 | 5 | 3 | 15 | 24 | 2 | 6 | 9 | 10 | 5 | 7 | 16 | 13 | 1 | 1 | 10 | 1 | 23 | 42 | 35 | 96 |
|  | Phi | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – |
|  | Qwen | 21 | 8 | 9 | 29 | 74 | 74 | 74 | 74 | 74 | 74 | 74 | 74 | 74 | 74 | 74 | 53 | 7 | 74 | 74 | 74 | 74 | 88 | 43 | 30 | 95 |
|  | GPT | **12** | 8 | 6 | 29 | 57 | 3 | 8 | **N/A** | 74 | 2 | 6 | 97 | 45 | 3 | 7 | 16 | 73 | 74 | **N/A** | **N/A** | 5 | 93 | 42 | 35 | 96 |
|  | **Gold** | **12** | **8** | **6** | **29** | **57** | **5** | **3** | **15** | **74** | **2** | **6** | **97** | **45** | **5** | **7** | **16** | **73** | **N/A** | **N/A** | **N/A** | **N/A** | **26** | **42** | **35** | **96** |
| Tritanopia | Llama | 19 | 2 | 5 | 17 | 1 | 3 | 2 | 4 | 1 | 3 | 4 | 42 | 46 | 4 | **7** | 3 | 8 | 4 | 4 | 4 | 1 | 3 | 4 | 3 | 8 |
|  | LLaVA | **12** | **8** | **6** | **29** | 37 | 3 | **3** | 16 | 7 | **2** | **6** | 9 | 1 | 6 | **7** | **16** | 3 | 1 | 1 | 1 | 12 | 28 | **42** | 38 | **96** |
|  | mPLUG | **12** | **8** | **6** | **29** | 37 | **5** | **3** | **15** | 24 | 9 | **6** | 9 | 10 | **5** | **7** | **16** | 13 | 10 | 10 | 10 | 10 | 23 | **42** | **35** | **96** |
|  | Phi | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – |
|  | Qwen | 21 | **8** | **6** | **29** | 37 | 74 | 74 | 74 | **74** | 74 | 74 | 74 | 74 | **5** | 3 | 53 | 3 | 74 | 0 | 74 | 74 | 38 | 43 | 30 | 95 |
|  | GPT | **12** | **8** | **6** | **29** | **57** | **N/A** | **3** | **15** | **74** | **2** | 3 | 37 | 36 | **5** | **7** | **16** | 33 | 74 | 74 | 6 | 3 | **N/A** | **42** | **35** | **96** |

Table 5: Predicted outputs of six LVLMs on Ishihara color-vision test plates under four simulated conditions (Normal, Protanopia, Deuteranopia, and Tritanopia) using the Base prompt only. The "Gold" row shows the ground-truth digits. Plates marked as "N/A" indicate those without a correct digit (e.g., blank or pattern-only plates) and are excluded from accuracy calculations.

strings, or collapse to a single digit, e.g., "70." Thus, strong Normal accuracy does not imply successful CVD simulation. Even with Linguistic or Visual Support, accuracy improves little and sometimes introduces new errors, indicating that prompt modifications alone cannot reproduce perceptual variation. Tritanopia shares the same ground-truth digits as Normal and therefore serves as a control condition. Models output nearly identical digits for both, matching clinical answers but indicating fallback to Normal rather than simulation. Severe failures under Protanopia and Deuteranopia further show that agreement in Tritanopia does not imply perceptual simulation.

*At the generation level*, current LVLMs assume normal color vision and, even with prompt support, fail to generate text that matches CVDs perception.

### 4.2 RQ2: Confidence Level

To evaluate model confidence, we measure *per-token perplexity* for each forced-decoded output. Comparing perplexity across conditions reveals whether confidence aligns with accuracy, whether linguistic or visual cues reduce uncertainty, and how strongly models default to normal vision.

Table 6 reports the mean and standard deviation (SD) of perplexity (shown as mean ± SD) for all models under the three prompt settings. Across the board, Normal and Tritanopia remain lower than Protanopia and Deuteranopia, indicating a pronounced bias toward normal vision, though exceptions are observed in Llama-3.2. One model rises from about $4 \pm 7$ on Normal to $70 \pm 80$ on Protanopia, while another climbs from about $14 \pm 13$ to $140 \pm 330$. Perplexity generally increases under Linguistic and Visual Support. Although visual cues lower perplexity for a few models, the tendency persists: Normal is always lowest and the red and green deficiencies remain highest.

Figure 2 plots perplexity for mPLUG-Owl3. We selected this model because it supports multiple images and exhibits the most stable perplexity across prompt settings. The three prompt settings share nearly identical shapes; Normal clusters near zero, whereas Protanopia and Deuteranopia rise to tall peaks. Adding a brief description or a reference image leaves the distribution almost unchanged, which shows that prompt tweaks do not alter the

| LVLM | Size | Prompt | Normal |  | Protanopia |  | Deuteranopia |  | Tritanopia |  |
|---|---|---|---|---|---|---|---|---|---|---|
|  |  |  | Mean | SD | Mean | SD | Mean | SD | Mean | SD |
| mPLUG-Owl3 | 7B | Base | 3.8 | 6.7 | 69.7 | 78.8 | 69.0 | 116.0 | 3.7 | 6.3 |
|  |  | Linguistic | 3.8 | 6.7 | 78.9 | 101.5 | 105.9 | 185.6 | 4.2 | 6.9 |
|  |  | Visual | 16.0 | 40.2 | 36.5 | 43.3 | 37.1 | 59.4 | 14.3 | 33.5 |
| Phi-3.5 | 4.2B | Base | 14.5 | 13.3 | 140.5 | 325.5 | 80.0 | 256.1 | 23.9 | 28.6 |
|  |  | Linguistic | 14.5 | 13.3 | 334.0 | 792.7 | 158.2 | 544.1 | 24.9 | 28.7 |
|  |  | Visual | 7.0 | 6.2 | 13.1 | 12.3 | 12.7 | 12.8 | 7.6 | 7.3 |
| Qwen2.5-VL | 7B | Base | 46.0 | 83.3 | 72.8 | 62.1 | 57.2 | 69.1 | 27.4 | 36.6 |
|  |  | Linguistic | 46.0 | 83.3 | 62.0 | 68.7 | 51.4 | 59.7 | 27.7 | 36.6 |
|  |  | Visual | 45.0 | 56.5 | 313.7 | 381.8 | 131.7 | 172.0 | 78.0 | 82.1 |

Table 6: Mean and standard deviation (SD) of *per-token perplexity* across four simulated vision conditions (Normal, Protanopia, Deuteranopia, Tritanopia) under three prompt settings (Base, Linguistic, Visual) for each LVLM. Only models that could be tested under all three settings are shown.

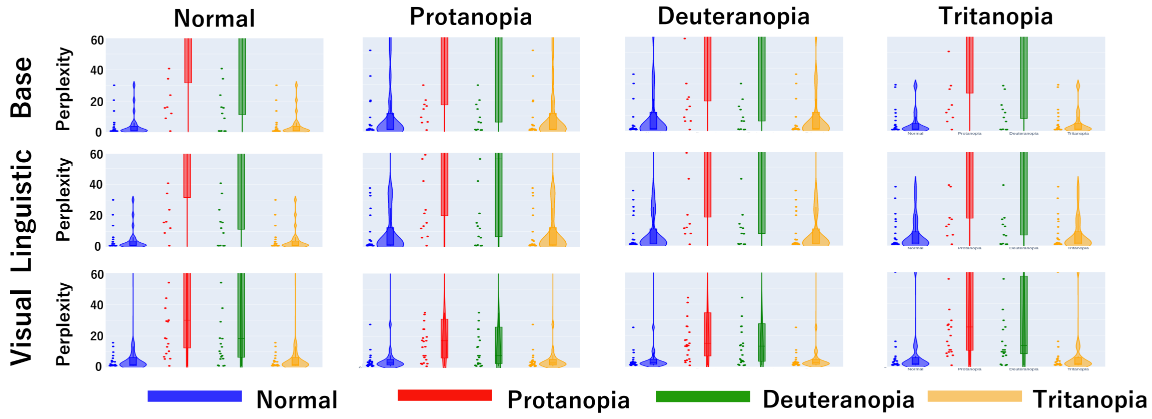


Figure 2: Violin plots of mPLUG-Owl3 perplexity on Ishihara digit prediction. Columns indicate the condition specified in the prompt (Normal, Protanopia, Deuteranopia, Tritanopia). Within each column, colors denote the vision type that defines the candidate digit whose perplexity is measured (Normal: blue, Protanopia: red, Deuteranopia: green, Tritanopia: yellow). Lower perplexity (for the gold answer) indicates higher model confidence.

| LVLM | Size | Normal | Protanopia |
|---|---|---|---|
| Llama-3.2 | 11B | 36.0 | 72.4 |
| LLaVA-NeXT | 13B | 20.0 | 86.2 |
| mPLUG-Owl3 | 7B | 28.0 | 86.2 |
| Phi-3.5 | 4.2B | 100.0 | 3.5 |
| Qwen2.5-VL | 7B | 4.0 | 86.2 |

Table 7: Doctor-style diagnosis accuracy (%) under Normal and Protanopia conditions for various LVLMs.

model's internal uncertainty. Other results are in Appendix E, showing the same trend: low perplexity for Normal vision, higher and variable for CVDs, with little effect from prompt changes.

*At the confidence level*, Normal vision consistently yields the lowest perplexity, while the CVD conditions result in similarly high perplexity, suggesting that models are fundamentally confused when simulating altered color vision and fail to reproduce such perceptual states.

### 4.3 RQ3: Internal Representation Level

We analyze whether LVLMs can perform doctor-style diagnosis by inferring vision types from responses using the Doctor-Style prompt (Table 3). This setting evaluates whether models can identify vision types from a response, analogous to how a doctor interprets Ishihara results. For each plate, we force-decode four labels (Normal, Protanopia, Deuteranopia, and Tritanopia) and select the most probable one. As shown in Table 7, models perform

| LVLM | Size | Normal | | | Protanopia | | | Deuteranopia | | | Tritanopia | | | Avg. | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Base | Ling. | Vis. | Base | Ling. | Vis. | Base | Ling. | Vis. | Base | Ling. | Vis. | Base | Ling. | Vis. |
| mPLUG-Owl3 | 7B | 71.4 | 71.4 | 62.0 | 5.9 | 5.9 | 5.9 | 5.9 | 5.9 | 17.7 | 66.7 | 66.7 | 62.0 | 37.5 | 37.5 | 36.9 |
| Qwen2.5-VL | 7B | 28.6 | 0.0 | 4.8 | 0.0 | 0.0 | 5.9 | 0.0 | 0.0 | 0.0 | 23.8 | 0.0 | 4.8 | 13.1 | 0.0 | 3.9 |
| GPT-4o | – | 90.5 | 90.5 | 90.5 | 23.6 | 17.7 | 23.6 | 5.9 | 23.5 | 5.9 | 71.4 | 61.9 | 42.9 | 47.9 | 48.4 | 40.7 |
| MedVLM-R1 | – | 57.1 | 57.1 | 0.0 | 5.9 | 5.9 | 0.0 | 5.9 | 5.9 | 0.0 | 61.9 | 61.9 | 0.0 | 32.7 | 32.7 | 0.0 |
| Qwen-DotPattern | – | 71.4 | 71.4 | 23.8 | 5.9 | 5.9 | 5.9 | 5.9 | 5.9 | 17.7 | 66.7 | 66.6 | 23.8 | 37.5 | 37.5 | 17.8 |

Table 8: *Digit accuracy* (%) of each LVLM under simulated vision types: **Normal**, **Protanopia**, **Deuteranopia**, and **Tritanopia**, with support settings: **Base**, **Ling.**, and **Vis.** As for **Vis.**, it was used only for models supporting multi-image input. This table extends Table 4 by including additional results for MedVLM-R1 and Qwen-DotPattern.

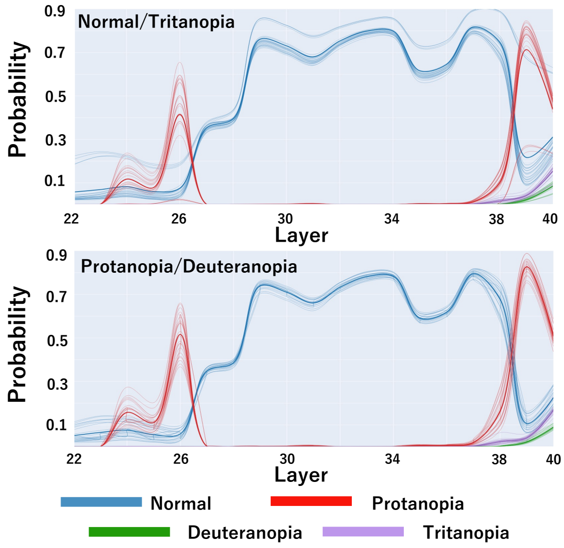


Figure 3: *Layer-wise diagnosis probabilities* for LLaVA, averaged over all plates.

well on Protanopia (over 70%, except Phi-3.5), but poorly on Normal (under 30%). This contrasts with Table 4, where they failed to recognize digits under red and green conditions.

To further examine this behavior, we use *Layer-Wise Diagnosis Probability*, inspired by Logit Lens (nostalgebraist, 2020), to track token-level probabilities across transformer layers. Figure 3 uses LLaVA-NeXT as a representative example. Results for other models are provided in Appendix F, where no consistent trends were observed across models, indicating that the layer-wise behavior varies depending on the architecture. LLaVA-NeXT tends to interpret images as Normal in mid-layers, shifting to Protanopia at the final layer, suggesting inconsistent processing. Nearly identical curves across ground-truths imply poor differentiation between conditions.

*At the internal representation level*, the model exhibits nearly identical behavior across all color vision types, showing no evidence of reasoning that reflects differences between conditions. This indicates a fundamental inability to understand and reproduce variations in color perception.

### 4.4 Domain Knowledge and Pattern Memorization

To deepen our analysis, we conducted two focused experiments as outlined below.

**Effect of medical expertise** As Ishihara plates originate in clinical practice, we expect that an LVLM trained on medical images and reports would outperform standard models. We therefore evaluated MedVLM-R1 (Pan et al., 2025), which is built on Qwen-2.0-VL (Wang et al., 2024) and fine-tuned on medical data, to investigate whether training on clinical data improves performance on color-based diagnostic tools.

**Dot pattern memorization hypothesis** A potential concern is that models may solve the task by exploiting superficial dot-pattern regularities in Ishihara plates, rather than demonstrating perceptual reasoning. To directly probe this, we fine-tuned Qwen2.0-VL on 10,000 synthetic Ishihara-style images released by Grover et al. (2025), producing a model specialized in dot-pattern recognition. We refer to this model as Qwen-DotPattern. This setup allows us to assess whether pattern memorization alone can explain model performance. Training details are provided in Appendix G.

Table 8 shows generation-level accuracy for MedVLM-R1 and Qwen-DotPattern. Under the Normal condition with the Base prompt, both outperform the Qwen2.5-VL baseline (MedVLM-R1: 57.1%, Qwen-DotPattern: 71.4%), indicating that domain-specific training with dot-pattern stimuli enhances recognition under typical vision. However, under Protanopia and Deuteranopia, follow

| LVLM | Size | Prompt | Normal | | Protanopia | | Deuteranopia | | Tritanopia | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Mean | SD | Mean | SD | Mean | SD | Mean | SD |
| Qwen2.5-VL | 7B | Base | 46.0 | 83.3 | 72.8 | 62.1 | 57.2 | 69.1 | 27.4 | 36.6 |
| | | Linguistic | 46.0 | 83.3 | 62.0 | 68.7 | 51.4 | 59.7 | 27.7 | 36.6 |
| | | Visual | 45.0 | 56.5 | 313.7 | 381.8 | 131.7 | 172.0 | 78.0 | 82.1 |
| MedVLM-R1 | 2B | Base | 15.5 | 28.6 | 166.4 | 325.3 | 86.1 | 160.8 | 14.7 | 26.2 |
| | | Linguistic | 15.5 | 28.6 | 189.7 | 369.6 | 91.5 | 195.2 | 13.2 | 24.5 |
| | | Visual | 48.6 | 63.6 | 83.7 | 92.3 | 72.7 | 89.6 | 50.9 | 62.1 |
| Qwen-DotPattern | 2B | Base | 4.7 | 8.9 | 79.2 | 136.3 | 75.4 | 171.7 | 4.7 | 6.6 |
| | | Linguistic | 4.7 | 8.9 | 106.3 | 213.0 | 100.2 | 249.9 | 4.7 | 6.5 |
| | | Visual | 6.2 | 9.1 | 41.5 | 55.2 | 28.7 | 47.1 | 7.4 | 9.9 |

Table 9: Mean and standard deviation (SD) of *per-token perplexity* across four simulated vision conditions under three prompt settings (Base, Linguistic, Visual). This table extends Table 6 by including additional results for MedVLM-R1 and Qwen-DotPattern.

| LVLM | Size | Normal | Protanopia |
|---|---|---|---|
| Qwen2.5-VL | 7B | 4.0 | 86.2 |
| MedVLM-R1 | 7B | 21.4 | 88.0 |
| Qwen-DotPattern | 7B | 85.7 | 0.0 |

Table 10: Doctor-style diagnosis accuracy (%) under Normal and Protanopia conditions for MedVLM-R1 and Qwen-DotPattern, extending Table 7.

the same trend as other models, showing low accuracy across prompts, suggesting that neither medical knowledge nor pattern memorization reproduces CVD perceptual distortions.

Table 9 presents the confidence level results in terms of perplexity. Under the Normal condition, both MedVLM-R1 and Qwen-DotPattern show lower perplexity than Qwen2.5-VL, indicating that domain adaptation enhances confidence on typical inputs. However, under red–green CVDs, both show higher perplexity than the base model, suggesting that medical or synthetic training improves confidence in normal vision but increases miscalibration when color perception is impaired.

We extend the doctor diagnosis analysis by adding MedVLM-R1 and Qwen-DotPattern, as shown in Table 10. MedVLM-R1 shows similar accuracy in both Normal (21.4%) and Protanopia (88.0%) settings, consistent with general model trends. This suggests that while medical training supports classification under impaired vision, it does not improve Normal vision. Qwen-DotPattern shows the opposite pattern. It performs well under Normal (85.7%) but fails under Protanopia (0.0 %). This suggests the model may rely on memorized dot layouts rather than learning generalizable structure, improving recognition in standard settings but reducing adaptability under altered vision.

These results show that training on domain-specific and dot-pattern images can improve model behavior in normal color vision, but do not lead to gains in reproducing the perceptual experience of CVDs. This highlights the value of the Ishihara Test, which remains a strong diagnostic benchmark despite its small size, as it is unaffected by pattern memorization and reliably evaluates a model's reasoning under impaired color perception.

## 5 Conclusion

This work examined whether LVLMs can simulate variation in human color perception using the Ishihara Test. We analyzed model behavior at the levels of generation, confidence, and internal representation to assess whether linguistic knowledge about CVDs reflects perceptual grounding. Although LVLMs can describe CVDs and the test procedure, they default to a normative color percept and do not reproduce the digit responses associated with different vision types. Additional textual guidance or visual manipulation does not induce alternative percepts, and probing reveals only weak separation between vision types in internal activations. These findings indicate that current LVLMs lack mechanisms for alternative perceptual experiences, highlighting the need for approaches that support perceptual diversity in future multimodal systems for accessible and human-centered deployment.

## 6 Limitations

**Dataset Scope and Scale.** Our evaluation relies on a small set of Ishihara plates that primarily target red–green color-vision deficiencies. The limited number and narrow focus reduce the diversity of conditions tested, so our findings may not generalise to other CVD types or broader visual tasks. While scaling to broader datasets and real-world tasks remains important future work, our focus here is on establishing a methodology and analysing behavioural differences across prompted vision conditions rather than memorisation. Possible pre-training exposure to Ishihara plates would mainly affect the Normal condition and does not account for failures under CVD prompts.

**Limitations in Evaluating Tritanopia.** The Ishihara Test is primarily designed to detect red–green deficiencies (Protanopia and Deuteranopia) and is not suitable for assessing blue–yellow deficiencies such as Tritanopia. Moreover, the medically defined ground-truth answers for Tritanopia coincide with those for Normal vision. In our setting, Tritanopia therefore functions largely as a control for instruction following rather than as a genuine benchmark for blue–yellow deficiencies. As a result, conclusions on Tritanopia are limited, and future work should evaluate all major CVD types using more general tasks beyond Ishihara plates.

**Evaluation Methodology.** We prompt LVLMs to simulate color-blind perception in a controlled diagnostic setting, tasking them with identifying digits in Ishihara plates. We use a restricted set of prompts (role-playing, doctor-perspective, step-by-step). While failures were broadly consistent across them, the prompt space is incomplete and other formulations may elicit different behaviours. Consequently, our conclusions characterise LVLM behaviour under this evaluation format rather than ruling out prompts that could induce more faithful CVD simulation. We also treat medically defined Ishihara digits as ground truth without validation from individuals with CVD or clinicians, which may limit ecological validity.

**Generalisability of Model Behavior.** Performance on stylised Ishihara plates does not imply that a model can handle color perception in broader settings. Models may rely on dot-pattern heuristics or memorised cues rather than simulating color confusion. Prior work also shows that vision–language models can exhibit color biases (Raj et al., 2024), suggesting that controlled diagnostic performance may not translate to broader color understanding. Caution is therefore required when extrapolating to natural images. Finally, this study is diagnostic rather than prescriptive: we do not test mitigation strategies, and developing such methods remains an important direction for CVD-aware LVLMs.

## 7 Ethical Considerations

**Inclusive Design for Perceptual Diversity.** Recent AI research communities, including ACL, have emphasized the importance of designing fair and inclusive systems that accommodate users with perceptual or cognitive differences. Color vision deficiency, which affects a significant portion of the global population, is one such factor that is often overlooked in model development and evaluation. By explicitly testing LVLM behavior under CVD conditions, this study contributes to a broader understanding of accessibility in multimodal AI systems and encourages further research in this direction. Building on this motivation, we also discuss fairness-related concerns in the context of model evaluation.

**Not for Diagnostic Use.** This study aims to evaluate whether models can simulate aspects of color vision deficiency as a step toward developing systems that better capture the diversity of human perception. It is not intended to replace professional diagnosis or support clinical decision-making. Even if a model produces correct responses to certain vision tests, these outputs must not be used for medical purposes or institutional assessments. To prevent such misuse, we clearly define the scope and limitations of our evaluation and explicitly state that the outputs are not a substitute for expert judgment in clinical contexts.

**Licensing and Copyright.** The Ishihara Test (Ishihara, 1917) is a long-established diagnostic tool widely used in academic research across fields such as ophthalmology, psychology, education, and machine learning (Zhao et al., 2019; Klinke et al., 2024; Lin et al., 2019). We did not reproduce, modify, or redistribute the plates; they were used solely for research and analysis.

**AI-assistant Usage.** Portions of the manuscript, such as prompt templates and wording adjustments, were drafted with the assistance of GPT-4 to ensure linguistic clarity. All technical content, analysis scripts, and final decisions were made by us.

## References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, and 110 others. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Jennifer Birch. 2012. Worldwide prevalence of red-green color deficiency. *J. Opt. Soc. Am. A Opt. Image Sci. Vis.*, 29(3):313–320.

Jirayu Burapacheep, Ishan Gaur, Agam Bhatia, and Tristan Thrush. 2024. Colorswap: A color and word order dataset for multimodal evaluation. *Preprint*, arXiv:2402.04492.

Yung-Sheng Chen and Yu-Chang Hsu. 1995. Computer vision on a colour blindness plate. *Image and Vision Computing*, 13(6):463–478.

Kara J Emery and Michael A Webster. 2019. Individual differences and their implications for color perception. *Curr. Opin. Behav. Sci.*, 30:28–33.

A Fanlo Zarazaga, J Gutiérrez Vásquez, and V Pueyo Royo. 2019. Revisión de los principales test clínicos para evaluar la visión del color. *Arch. Soc. Esp. Oftalmol. (Engl. Ed.)*, 94(1):25–32.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Rynaa Grover, Jayant Sravan Tamarapalli, Sahiti Yerramilli, and Nilay Pande. 2025. Huemanity: Probing fine-grained visual perception in mllms. *Preprint*, arXiv:2506.03194.

Yu Hao, Fan Yang, Hao Huang, Shuaihang Yuan, Sundeep Rangan, John-Ross Rizzo, Yao Wang, and Yi Fang. 2024. A multi-modal foundation model to assist people with blindness and low vision in environmental interaction. *Journal of Imaging*, 10(5).

Christopher Henshaw, Jacob Dennis, Jonathan Nadzam, and Alan J. Michaels. 2025. Number recognition through color distortion using convolutional neural networks. *Computers*, 14(2).

Shinobu Ishihara. 1917. *Tests for Colour-Blindness*. Handaya, Tokyo, Hongo Harukicho.

Genji Kawakita, Ariel Zeleznikow-Johnston, Naotsugu Tsuchiya, and Masafumi Oizumi. 2024. Gromov–wasserstein unsupervised alignment reveals structural correspondences between the color similarity structures of humans and large language models. *Scientific Reports*, 14(1):15917.

Corentin Kervadec, Grigory Antipov, Moez Baccouche, and Christian Wolf. 2021. Roses are red, violets are blue... but should vqa expect them to? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2776–2785.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5583–5594. PMLR.

Thomas Klinke, Wolfgang Hannak, Klaus Böning, and Holger Jakstat. 2024. A comparative study of the sensitivity and specificity of the ishihara test with various displays. *International Dental Journal*, 74(4):892–896. Epub 2024 Jan 15.

Yijun Liang, Ming Li, Chenrui Fan, Ziyue Li, Dang Nguyen, Kwesi Cobbina, Shweta Bhardwaj, Jiuhai Chen, Fuxiao Liu, and Tianyi Zhou. 2025. Colorbench: Can vlms see and understand the colorful world? a comprehensive benchmark for color perception, reasoning, and robustness. *Preprint*, arXiv:2504.10514.

Huei-Yung Lin, Li-Qi Chen, and Min-Liang Wang. 2019. Improving discrimination in color vision deficiency by image re-coloring. *Sensors*, 19(10).

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

Gustavo M. Machado, Manuel M. Oliveira, and Leandro A. F. Fernandes. 2009. A physiologically-based model for simulation of color vision deficiency. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1291–1298.

Martín Montes Rivera, Alejandro Padilla, Juana Canul-Reich, and Julio Ponce. 2020. *Realtime recoloring Ishihara plates using artificial neural networks for helping colorblind*.

Shogo Morita, Yan Zhang, Takuto Yamauchi, Sinan Chen, Jialong Li, and Kenji Tei. 2024. Towards context-aware support for color vision deficiency: An approach integrating llm and ar. *Preprint*, arXiv:2407.04362.

nostalgebraist. 2020. Interpreting gpt: the logit lens. https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens. LessWrong.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Madry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Cory Paik, Stéphane Aroca-Ouellette, Alessandro Roncone, and Katharina Kann. 2021. The World of an Octopus: How Reporting Bias Influences a Language Model's Perception of Color. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 823–835, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jiazhen Pan, Che Liu, Junde Wu, Fenglin Liu, Jiayuan Zhu, Hongwei Bran Li, Chen Chen, Cheng Ouyang, and Daniel Rueckert. 2025. Medvlm-r1: Incentivizing medical reasoning capability of vision-language models (vlms) via reinforcement learning. *Preprint*, arXiv:2502.19634.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Pooyan Rahmanzadehgervi, Logan Bolton, Mohammad Reza Taesiri, and Anh Totti Nguyen. 2024. Vision language models are blind. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, pages 18–34.

Chahat Raj, Anjishnu Mukherjee, Aylin Caliskan, Antonios Anastasopoulos, and Ziwei Zhu. 2024. BiasDora: Exploring hidden biased associations in vision-language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10439–10455, Miami, Florida, USA. Association for Computational Linguistics.

Debi Roberson and J Richard Hanley. 2007. Color vision: color categories vary with language after all. *Current Biology*, 17(15):R605–R607.

Ahnaf Mozib Samin, M Firoz Ahmed, and Md. Mushtaq Shahriyar Rafee. 2025. ColorFoil: Investigating color blindness in large vision and language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pages 294–300, Albuquerque, USA. Association for Computational Linguistics.

Lisa Schut, Yarin Gal, and Sebastian Farquhar. 2025. Do multilingual llms think in english? *arXiv preprint arXiv:2502.15603*.

Haz Sameen Shahgir, Khondker Salman Sayeed, Abhik Bhattacharjee, Wasi Uddin Ahmad, Yue Dong, and Rifat Shahriyar. 2024. Illusionvqa: A challenging optical illusion dataset for vision language models. *Preprint*, arXiv:2403.15952.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in English? on the latent language of multilingual transformers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15366–15394, Bangkok, Thailand. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Lingling Xu, Haoran Xie, Si-Zhao Joe Qin, Xiaohui Tao, and Fu Lee Wang. 2023. Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment. *Preprint*, arXiv:2312.12148.

Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2024. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. *Preprint*, arXiv:2408.04840.

Jiawei Zhao, Michael Joseph Fliotsos, Mehrnaz Ighani, and Allen O Eghrari. 2019. Comparison of a smartphone application with ishihara pseudoisochromatic plate for testing colour vision. *Neuro-ophthalmology*, 43(4):235–239. Epub 2018 Nov 19.

Zewei Zhou, Tianhui Cai, Seth Z. Zhao, Yun Zhang, Zhiyu Huang, Bolei Zhou, and Jiaqi Ma. 2025. AutoVLA: A vision-language-action model for end-to-end autonomous driving with adaptive reasoning and reinforcement fine-tuning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, Quan Vuong, Vincent Vanhoucke, Huong Tran, Radu Soricut, Anikait Singh, Jaspiar Singh, Pierre Sermanet, Pannag R. Sanketi, Grecia Salazar, and 35 others. 2023. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Proceedings of The 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pages 2165–2183. PMLR.

## A Detailed Model Settings

In this study, to ensure fair and consistent performance comparisons, all experiments were conducted using a single NVIDIA RTX 6000 Ada GPU. All models were run under the same computational settings, with generation performed using half quantization to reduce memory usage while maintaining model fidelity. We used publicly available models from Hugging Face and OpenAI, as summarized in Table 11. Generation was conducted using greedy decoding, with fixed random seeds to ensure reproducibility. All other settings adhered to the default configurations of each model implementation.

| Model | Params | HuggingFace Name / OpenAI Name |
|---|---|---|
| Phi3.5 | 4.2B | microsoft/Phi-3-vision-128k-instruct |
| Qwen2.5-VL | 7B | Qwen/Qwen-2.5-VL-7B-Instruct |
| mPLUG | 7B | mPLUG/mPLUG-Owl3-7B-240728 |
| Llama3.2 | 11B | meta-llama/Llama-3.2-11B-Vision-Instruct |
| LLaVA-NeXT | 13B | llava-hf/llava-v1.6-vicuna-13b-hf |
| MedVLM-R1 | 2B | JZPeterPan/MedVLM-R1 |
| GPT-4o | – | gpt-4o-2024-11-20 |

Table 11: Detailed model names.

## B Language Knowledge of the Ishihara Test and CVDs

Before testing visual ability, we verified that all LVLMs possess basic language knowledge of the Ishihara Test and the three major color vision deficiencies. As the main paper includes only one example, this appendix provides the full outputs for the remaining models.

Tables 19–23 show representative answers from Llama-3.2, LLaVA-NeXT, mPLUG-Owl3, Phi-3.5, and Qwen2.5-VL. Each table has input prompts on the left and trimmed model responses on the right. All models except Phi-3.5, which returned an error for the Ishihara query, accurately describe the test and define Protanopia, Deuteranopia, and Tritanopia, confirming sufficient textual knowledge. These results show that failures in the main experiments arise from weaknesses in visual grounding, not language understanding.

**Prompts in Detail** For each color vision condition, we designed three types of prompts: Base, Linguistic Support, and Visual Support. The full prompt templates are provided in Tables 14–16.

Table 14 presents the *Base* prompts, which specify only the condition (e.g., “You are Protanopic”) and then ask the model to report the number shown in the Ishihara plate. This represents the simplest setting, testing whether the model can directly simulate the specified condition.

Table 15 illustrates the *Linguistic Support* prompts, where a short textual description of the impairment (e.g., “red tones appear dimmer and may be confused with green or brown”) is added. This design evaluates whether providing linguistic context improves the model’ s responses.

Table 16 shows the *Visual Support* prompts. In this case, the condition is accompanied by a few shot reference example (e.g., “A Protanopic person sees this number: {few_shot_number}”) before asking the model to identify the number in a new plate. This setup tests whether visual exemplars serve as useful guidance for simulating CVD perception.

Together, these tables make explicit how the three prompt types were implemented, complementing the overview provided in Section 3.3.

## C Effects of Prompt Variants on Simulating CVD Perception

Prompt phrasing can influence how LVLMs interpret images, particularly when the task requires aligning visual input with an altered perceptual perspective. To assess whether failures in our main experiments were attributable to prompt design rather than perceptual grounding, we evaluated four alternative prompt formulations for the Protanopia condition:

- **Prompt 1 (Description-Based)**: “What number would a Protanopic viewer likely see in this image? Only provide the predicted number."

- **Prompt 2 (Doctor Perspective)**: "You are an eye doctor evaluating how different color vision types perceive this test plate. What number would a Protanopic patient report? Only output the number."
- **Prompt 3 (Role-Play Perspective)**: "Answer as if you are a Protanopic viewer. Respond with the number you visually perceive. Only output the number."
- **Prompt 4 (Step-by-Step Reasoning)**: "Step 1: Describe how the red and green dots differ under Protanopia. Step 2: Based on Step 1, what number appears to a Protanopic viewer? Only output Step 2."

Table 12 reports digit accuracy under Normal and three CVD conditions for Prompts 1–3. Prompt 4 was excluded due to highly unstable behavior: models frequently produced free-form explanations, failed to emit a digit token in Step 2, or ignored the requested output format entirely.

Across all prompt variants, we observe moderate variability in the Normal condition, confirming that prompt phrasing can influence how LVLMs interpret visual input when no perceptual transformation is required. However, under Protanopia, Deuteranopia, and Tritanopia, accuracy remains low across prompts and models. Neither altering task framing (doctor vs. viewer) nor introducing explicit role-playing improved condition-appropriate responses. These results indicate that prompt formulation alone is insufficient to induce LVLMs to simulate condition-specific perceptual distortions, consistent with the main conclusion that limitations arise from insufficient visual grounding rather than linguistic control alone.

## D Digit Results: Detailed Analysis

Tables 17 and 18 report the raw digit outputs that each LVLM produced for Ishihara plates under four simulated conditions (Normal, Protanopia, Deuteranopia, Tritanopia). Unless stated otherwise, plates marked as "N/A" (no correct digit) are excluded from accuracy and are discussed only for error behavior.

### D.1 Linguistic Support prompts (Table 17).

Adding a short textual description yields only modest and inconsistent gains. Under Normal vision, GPT remains strong and LLaVA/mPLUG perform similarly to the Base setting, while Llama continues to miss many plates. Under simulated CVDs, condition-specific digits are seldom produced; the same plates that were difficult under Base remain challenging, and response biases (e.g., repeated 74/42) persist. Hallucinated digits also appear on N/A plates, showing that brief definitions alone are insufficient to induce condition-aware perception.

### D.2 Visual Support prompts (Table 18).

Providing a few shot reference image before the target plate produces mixed results. GPT maintains strong accuracy under Normal vision but shows limited and inconsistent improvements for simulated CVDs. mPLUG performs slightly better on some plates compared to Base/Linguistic settings but still fails to reliably follow the instructed condition. Phi occasionally outputs correct answers without cross-plate consistency, while Qwen remains unstable and frequently repeats the same digits. Some regressions and hallucinated digits are also observed, indicating that single visual exemplars do not reliably induce condition-appropriate perception.

### D.3 Overall Findings

Across all prompt types, LVLMs display recurring error patterns: frequent mode collapse to a small set of digits (e.g., 74, 42, 3), insensitivity to the instructed CVD condition, and hallucinated outputs on plates without valid digits. These behaviors show that neither short textual descriptions nor single visual exemplars are sufficient to ground condition-specific perception. While the models can describe CVDs accurately in text, they fail to translate these instructions into consistent digit recognition on Ishihara plates. This confirms that the main limitation lies in visual grounding rather than language knowledge or prompt design.

## E Additional Perplexity Analyses

In the main text, we focused on mPLUG-Owl3 (Figure 2) because it supports multi-image input and shows the most stable perplexity across prompt settings. For completeness, we report violin plots of perplexity distributions for three additional models: LLaMA, Phi, and Qwen.

Figures 4–6 show perplexity measured by force decoding the Gold Answer for Ishihara digit predictions under four simulated vision conditions (Normal, Protanopia, Deuteranopia, Tritanopia) and three prompt types (Base, Linguistic, Visual).

| LVLM | Size | Normal | | | Protanopia | | | Deuteranopia | | | Tritanopia | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **P1** | **P2** | **P3** | **P1** | **P2** | **P3** | **P1** | **P2** | **P3** | **P1** | **P2** | **P3** |
| Llama-3.2 | 11B | 9.5 | 4.8 | 9.5 | 17.6 | 11.8 | 0.0 | 11.8 | 11.8 | 11.8 | 4.8 | 4.8 | 4.8 |
| LLaVA-NeXT | 13B | 61.9 | 52.4 | 57.1 | 5.9 | 5.9 | 5.9 | 5.9 | 5.9 | 5.9 | 47.6 | 61.9 | 57.1 |
| mPLUG-Owl3 | 7B | 71.4 | 76.2 | 66.7 | 5.9 | 5.9 | 5.9 | 5.9 | 5.9 | 5.9 | 66.7 | 71.4 | 66.7 |
| Phi-3.5 | 4.2B | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| Qwen2.5-VL | 7B | 28.6 | 28.6 | 38.1 | 5.9 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 14.3 | 19.1 | 28.6 |

Table 12: **Digit accuracy** (%) under three prompt variants (**P1–P3**) across simulated vision types. Prompt 4 is omitted due to unstable output formatting and inconsistent digit emission.

Across these models, we observe similar tendencies as discussed in the main text: Normal vision yields consistently lower perplexity, while the CVD conditions lead to higher and more variable perplexity. Adding textual descriptions (Linguistic Support) or reference images (Visual Support) does not substantially alter these distributions. This supports the conclusion that prompt modifications are insufficient to reduce model uncertainty when simulating altered color vision.

## F Detailed Diagnosis of CVDs with LVLMs

Figures 7–10 extend the main analysis by visualizing token-level label probabilities across transformer layers for four additional LVLMs. Each panel groups plates by the correct vision condition and plots plate-wise trajectories (thin lines) together with the average curve (bold line). Left plots correspond to Normal or Tritanopia, right plots to Protanopia or Deuteranopia.

**Llama-3.2 (Figure 7)**, For Normal/Tritanopia, the Normal label stays near zero until three layers before the head, then rises above 50%. Red and green show an analogous late surge in the Protanopia curve, reflecting a rise in uncertainty.

**mPLUG-Owl3 (Figure 8)**, Early layers are mixed but converge rapidly. For instance, tritanopia exceeds 80% by layer 26. When the digit is unreadable, the model shifts probability toward Tritanopia, possibly explaining its low variance under Base.

**Phi-3.5 (Figure 9)**, Curves remain diffuse and never exceed 25% even near the output. This matches the wide perplexity spread and frequent no-answer cases.

**Qwen2.5-VL (Figure 10)**, Normal shows local peaks rather than a smooth ascent. Red and green plates present competing red and green waves that collapse to a single token at the final layer. However, the overall probabilities remain extremely low, on the order of $10^{-6}$ even beyond layer 6, indicating very weak model confidence.

## G Fine-tuning and Inference Settings

| Hyperparameter | Value |
|---|---|
| Model | Qwen/Qwen2-VL-2B-Instruct |
| Torch dtype | bfloat16 |
| Seed | 42 |
| Max input length | 2048 |
| Epochs | 1 |
| Learning rate | 2e-4 |
| Batch size | 1 |
| Gradient accumulation | 32 |
| Warmup ratio | 0 (not used) |
| Gradient checkpointing | True |
| LoRA rank ($r$) | 8 |
| LoRA alpha | 16 |
| LoRA dropout | 0.05 |
| Save steps | 200 |
| Eval strategy | None |

Table 13: Hyperparameters and settings used for fine-tuning. We used `Transformers` (Wolf et al., 2020), `PEFT` (Xu et al., 2023), and `trl` libraries with default configurations unless otherwise noted.

| Condition | Prompt |
|---|---|
| **Protanopia** | `You are Protanopic.`<br>`What number do you see?`<br>`Only provide the answer. If you see nothing, answer 'Nothing'. No additional information.` |
| **Deuteranopia** | `You are Deuteranopic.`<br>`What number do you see?`<br>`Only provide the answer. If you see nothing, answer 'Nothing'. No additional information.` |
| **Tritanopia** | `You are Tritanopic.`<br>`What number do you see?`<br>`Only provide the answer. If you see nothing, answer 'Nothing'. No additional information.` |
| **Normal Vision** | `What number do you see?`<br>`Only provide the answer. If you see nothing, answer 'Nothing'. No additional information.` |

Table 14: Base prompts for each color-vision condition.

| Setting | Prompt |
|---|---|
| **Protanopia** | `You are Protanopic.`<br>`Protanopia means you have difficulty distinguishing red tones, as red appears dimmer and may be confused with green or brown.`<br>`What number do you see?`<br>`Only provide the answer. If you see nothing, answer 'Nothing'. No additional information.` |
| **Deuteranopia** | `You are Deuteranopic.`<br>`Deuteranopia means you have difficulty distinguishing green tones, as green appears dimmer and may be confused with red or brown.`<br>`What number do you see?`<br>`Only provide the answer. If you see nothing, answer 'Nothing'. No additional information.` |
| **Tritanopia** | `You are Tritanopic.`<br>`Tritanopia means you have difficulty distinguishing blue tones, as blue appears dimmer and may be confused with green or gray.`<br>`What number do you see?`<br>`Only provide the answer. If you see nothing, answer 'Nothing'. No additional information.` |
| **Normal Vision** | `What number do you see?`<br>`Only provide the answer. If you see nothing, answer 'Nothing'. No additional information.` |

Table 15: Prompts with linguistic support for each color-vision condition.

| Condition | Prompt |
|---|---|
| **Protanopia** | You are Protanopic.<br>You have difficulty distinguishing certain colors.<br>Example 1:<br>Image shows an Ishihara plate.<br>A Protanopic person sees this number: {few_shot_number}<br>Now, look at the following image.<br>What number do you see?<br>Only provide the answer. If you see nothing, answer ’Nothing’. No additional information. |
| **Deuteranopia** | You are Deuteranopic.<br>You have difficulty distinguishing certain colors.<br>Example 1:<br>Image shows an Ishihara plate.<br>A Deuteranopic person sees this number: {few_shot_number}<br>Now, look at the following image.<br>What number do you see?<br>Only provide the answer. If you see nothing, answer ’Nothing’. No additional information. |
| **Tritanopia** | You are Tritanopic.<br>You have difficulty distinguishing certain colors.<br>Example 1:<br>Image shows an Ishihara plate.<br>A Tritanopic person sees this number: {few_shot_number}<br>Now, look at the following image.<br>What number do you see?<br>Only provide the answer. If you see nothing, answer ’Nothing’. No additional information. |
| **Normal Vision** | You have normal color vision.<br>Example 1:<br>Image shows an Ishihara plate.<br>A person with normal color vision sees this number: {few_shot_number}<br>Now, look at the following image.<br>What number do you see?<br>Only provide the answer. If you see nothing, answer ’Nothing’. No additional information. |

Table 16: Prompts with visual support examples for each color-vision condition.

| | | | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Gold** | **12** | **8** | **6** | **29** | **57** | **5** | **3** | **15** | **74** | **2** | **6** | **97** | **45** | **5** | **7** | **16** | **73** | **N/A** | **N/A** | **N/A** | **N/A** | **26** | **42** | **35** | **96** |
| | LLama | 16 | **8** | 8 | 42 | 52 | 4 | 4 | 52 | 15 | 3 | 15 | 42 | 3 | 4 | **7** | 4 | 15 | 15 | 4 | 15 | 6 | 3 | 4 | 52 | 52 |
| Normal | LLava | **12** | **8** | **6** | **29** | 37 | 3 | 6 | 16 | 7 | **2** | **6** | 9 | 5 | 7 | **7** | **16** | 3 | 1 | 1 | 1 | 1 | 28 | **42** | **35** | **96** |
| | mPlug | **12** | **8** | **6** | **29** | 37 | **5** | **3** | **15** | 24 | **2** | **6** | 9 | 10 | **5** | **7** | **16** | 13 | 100 | 100 | 10 | 100 | 23 | **42** | **35** | **96** |
| | Phi | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – |
| | Qwen | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – |
| | GPT | **12** | **8** | **6** | **29** | **57** | **5** | **3** | **15** | **74** | **2** | **6** | 74 | **45** | **5** | **7** | **16** | 23 | 26 | 74 | 29 | 5 | **26** | **42** | **35** | **96** |
| | **Gold** | **12** | **3** | **5** | **70** | **35** | **2** | **5** | **17** | **21** | **N/A** | **N/A** | **N/A** | **N/A** | **N/A** | **N/A** | **N/A** | **N/A** | **5** | **2** | **45** | **73** | **6** | **2** | **5** | **6** |
| | LLama | 15 | 8 | 7 | 42 | 5 | 6 | 2 | 3 | 9 | 8 | 4 | 2 | 3 | 4 | 4 | 3 | 6 | 2 | 6 | 4 | 4 | **6** | 52 | 2 | 42 |
| Protanopia | LLava | **12** | 8 | 6 | 29 | 37 | 3 | 6 | 16 | 7 | 2 | 6 | 9 | 1 | 6 | 7 | 16 | 3 | 1 | 1 | 1 | 1 | 26 | 42 | 38 | 96 |
| | mPlug | **12** | 8 | 6 | 29 | 37 | 5 | 3 | 15 | 24 | 9 | 6 | 9 | 10 | 5 | 7 | 16 | 13 | 1 | 1 | 9 | 1 | 23 | 42 | 35 | 96 |
| | Phi | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – |
| | Qwen | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – |
| | GPT | 32 | **3** | **5** | 26 | 5 | 8 | 8 | 13 | **21** | 5 | 3 | 37 | 46 | 2 | 7 | 16 | 13 | 3 | 5 | 5 | 5 | 23 | 12 | 30 | 39 |
| | **Gold** | **12** | **3** | **5** | **70** | **35** | **2** | **5** | **17** | **21** | **N/A** | **N/A** | **N/A** | **N/A** | **N/A** | **N/A** | **N/A** | **N/A** | **5** | **2** | **45** | **73** | **2** | **4** | **3** | **9** |
| | LLama | 8 | 8 | 6 | 59 | 3 | 4 | 6 | 4 | 66 | 6 | 8 | 22 | 49 | 7 | 7 | 4 | 4 | 8 | 7 | 11 | 41 | **2** | 42 | 33 | 44 |
| Deuteranopia | LLava | **12** | 8 | 6 | 29 | 57 | 3 | 6 | 16 | 14 | 2 | 6 | 9 | 1 | 6 | 7 | 16 | 3 | 1 | 12 | 12 | 12 | 26 | 42 | 38 | 96 |
| | mPlug | **12** | 8 | 6 | 29 | 37 | 5 | 3 | 15 | 24 | 9 | 6 | 9 | 10 | 5 | 7 | 16 | 13 | 1 | 1 | 9 | 1 | 23 | 42 | 35 | 96 |
| | Phi | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – |
| | Qwen | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – |
| | GPT | **12** | **3** | **5** | 29 | 37 | **2** | 8 | 13 | 41 | 2 | 5 | 97 | 45 | 3 | 3 | 16 | 78 | 6 | 5 | 3 | 74 | 93 | 42 | 35 | 96 |
| | **Gold** | **12** | **8** | **6** | **29** | **57** | **5** | **3** | **15** | **74** | **2** | **6** | **97** | **45** | **5** | **7** | **16** | **73** | **N/A** | **N/A** | **N/A** | **N/A** | **26** | **42** | **35** | **96** |
| | LLama | **12** | 2 | 8 | 17 | 8 | 4 | **3** | 6 | 2 | **2** | 2 | 8 | 7 | **5** | **7** | 7 | 6 | 2 | 4 | 2 | 4 | 6 | 4 | 4 | 8 |
| Tritanopia | LLava | **12** | **8** | **6** | **29** | **57** | 3 | **3** | 16 | 7 | **2** | **6** | 9 | 1 | 6 | **7** | **16** | 3 | 1 | 1 | 12 | 12 | **26** | **42** | 38 | **96** |
| | mPlug | **12** | **8** | **6** | **29** | 37 | **5** | **3** | **15** | 24 | 9 | **6** | 9 | 10 | **5** | **7** | **16** | 13 | 10 | 10 | 10 | 10 | 23 | **42** | **35** | **96** |
| | Phi | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – |
| | Qwen | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – |
| | GPT | **12** | **8** | **6** | **29** | **57** | 2 | 8 | **15** | 41 | **2** | 3 | 73 | 35 | **5** | **7** | **16** | **73** | 3 | 5 | 5 | 3 | 93 | **42** | **35** | **96** |

Table 17: Predicted outputs of six LVLMs (Llama, Llava, mPlug, Phi, Qwen, and GPT) on Ishihara color-vision test plates under four simulated conditions, using Linguistic Support prompts that include a brief explanation of how each deficiency alters color perception. The "Gold" row shows the ground-truth digits.

| | | | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Gold** | **12** | **8** | **6** | **29** | **57** | **5** | **3** | **15** | **74** | **2** | **6** | **97** | **45** | **5** | **7** | **16** | **73** | **N/A** | **N/A** | **N/A** | **N/A** | **26** | **42** | **35** | **96** |
| | mPlug | **12** | **8** | **6** | **29** | 37 | **5** | **3** | **15** | 2 | 6 | 97 | 90 | 5 | **5** | **7** | **16** | 13 | N/A | N/A | N/A | N/A | **26** | **42** | **35** | **96** |
| Normal | Phi | **12** | 6 | 29 | 37 | 5 | 3 | **15** | 74 | **2** | **6** | **97** | **45** | **5** | **7** | **16** | 73 | 1 | 3 | 3 | 3 | 25 | **42** | 3 | 96 | 90 |
| | Qwen | 8 | 6 | 38 | 37 | 74 | 74 | 15 | 74 | 74 | 74 | 3 | 37 | 74 | 3 | 16 | **73** | 4 | 74 | 42 | 74 | 74 | 37 | 24 | 74 | 74 |
| | GPT | **12** | **8** | **6** | **29** | **57** | **5** | **3** | **15** | **74** | **2** | **6** | **97** | **45** | **5** | **7** | **16** | 33 | 74 | 74 | N/A | 74 | 93 | **42** | **35** | **96** |
| | **Gold** | **12** | **3** | **5** | **70** | **35** | **2** | **5** | **17** | **21** | **N/A** | **N/A** | **N/A** | **N/A** | **N/A** | **N/A** | **N/A** | **N/A** | **5** | **2** | **45** | **73** | **6** | **2** | **5** | **6** |
| | mPlug | **12** | 8 | 6 | 29 | 37 | 5 | 3 | 15 | 2 | 2 | 6 | 9 | 5 | 5 | 7 | 13 | 13 | 1 | **45** | **73** | **6** | **2** | 43 | 3 | **96** |
| Protanopia | Phi | **12** | 6 | **5** | 35 | 3 | 1 | 17 | 1 | 3 | 3 | 4 | 17 | 3 | 7 | 7 | 3 | 1 | **2** | **45** | **73** | 3 | 3 | 3 | 3 | 9 |
| | Qwen | 7 | **3** | 8 | 74 | 7 | 7 | 3 | 74 | N/A | 74 | 74 | 74 | 3 | 3 | 8 | N/A | 7 | 7 | 74 | 42 | 4 | 7 | 3 | 3 | 7 |
| | GPT | 21 | **3** | **5** | 26 | 5 | 5 | 5 | 15 | N/A | 2 | 3 | 37 | **45** | 5 | N/A | 16 | 3 | 5 | 74 | **6** | 3 | 25 | 42 | 33 | **96** |
| | **Gold** | **12** | **3** | **5** | **70** | **35** | **2** | **5** | **17** | **21** | **N/A** | **N/A** | **N/A** | **N/A** | **N/A** | **N/A** | **N/A** | **N/A** | **5** | **2** | **45** | **73** | **2** | **4** | **3** | **9** |
| | mPlug | **12** | 8 | 6 | 29 | 37 | 5 | 3 | 15 | 2 | 2 | 6 | 9 | 5 | 5 | 7 | 13 | 13 | 1 | **45** | **73** | **2** | 2 | 4 | 35 | **96** |
| Deuteranopia | Phi | **12** | 5 | | 35 | | | 17 | | | | | | | 7 | 7 | | | | **45** | **73** | | 3 | 3 | 9 | 9 |
| | Qwen | 7 | 7 | 3 | 74 | 7 | 7 | 23 | 74 | 74 | 74 | 74 | 74 | 74 | 74 | 8 | 74 | 7 | 7 | 74 | 42 | 7 | 7 | 7 | 7 | 7 |
| | GPT | 14 | 0 | **5** | 29 | 37 | 5 | 9 | 15 | 74 | 2 | 3 | 97 | **45** | 5 | N/A | 16 | 33 | 2 | 17 | 74 | 5 | 23 | 42 | **35** | 96 |
| | **Gold** | **12** | **8** | **6** | **29** | **57** | **5** | **3** | **15** | **74** | **2** | **6** | **97** | **45** | **5** | **7** | **16** | **73** | **N/A** | **N/A** | **N/A** | **N/A** | **26** | **42** | **35** | **96** |
| | mPlug | **12** | **8** | **6** | **29** | 37 | **5** | **3** | **15** | 2 | **6** | 6 | 90 | **5** | 5 | 7 | 13 | 3 | N/A | N/A | N/A | 26 | **42** | 42 | **35** | **96** |
| Tritanopia | Phi | **12** | 6 | | 23 | **5** | **3** | **15** | 74 | | **6** | 9 | **45** | | 7 | 7 | **73** | | | | | **26** | **42** | 3 | **96** | 9 |
| | Qwen | 3 | 3 | 3 | 3 | 4 | 7 | 15 | 3 | 4 | 3 | 3 | 3 | 3 | 3 | 1 | 42 | N/A | N/A | N/A | N/A | 4 | 3 | 42 | 3 | 3 |
| | GPT | 21 | 3 | **6** | **29** | N/A | **5** | 8 | **15** | 71 | **2** | 5 | 57 | 46 | 3 | 3 | **16** | N/A | N/A | N/A | N/A | 57 | 26 | **42** | **35** | 86 |

Table 18: Predicted outputs of four LVLMs (mPlug, Phi, Qwen, and GPT) on Ishihara color-vision test plates under four simulated conditions, using Visual Support prompts that present an illustrative example image of each plate as seen under the deficiency. The "Gold" row shows the ground-truth digits. Evaluation was performed only on models capable of ingesting two or more images in a single turn.

| Input | Model Output (LLama as an example) |
|---|---|
| What is Ishihara Test? | This image is used to test color vision. It is called an Ishihara plate, named after its creator, Dr. Shinobu Ishihara, a Japanese professor of ophthalmology. The plate is a widely used tool for diagnosing color blindness, particularly red-green color blindness, which is the most common form of color vision deficiency. The plate consists of a series of dots in different colors and sizes, arranged in a specific pattern. The viewer is asked to identify a number or shape hidden within the dots. For people with normal color vision, the correct answer is obvious, but for those with color vision deficiency, the correct answer may not be apparent. In this image, the number 8̈ïs hidden among the dots. The dots are arranged in a way that creates a red 8̈s̈hape, but the dots are also scattered randomly throughout the plate, making it difficult to see the number without looking closely. For someone with normal color vision, the red 8̈s̈tands out clearly, but for someone with red-green color blindness, the number may not be visible at all. The Ishihara plate is used to diagnose color blindness because it exploits the fact that people with certain types of color vision deficiency have difficulty distinguishing between certain colors. In this case, the plate uses the fact that people with red-green color blindness have trouble distinguishing between red and green colors. By using a combination of red and green dots, the plate can help identify individuals who have difficulty seeing these colors. Overall, the Ishihara plate is a simple yet effective tool for diagnosing color blindness, and it has been widely used in medical settings for many years. |
| What is Protanopia? | Protanopia, also known as red-blindness, is a type of color vision deficiency that affects approximately 1% of males and 0.05% of females. It is caused by a mutation in the OPN1LW gene, which codes for the long-wavelength cone cells in the retina. These cone cells are responsible for detecting long-wavelength light, which is associated with the color red. In individuals with protanopia, the long-wavelength cone cells are either absent or non-functional, leading to a reduced ability to detect red light. As a result, people with protanopia may have difficulty distinguishing between red |
| What is Deuteranopia? | Deuteranopia is a type of color vision deficiency that affects approximately 5% of males and 0.4% of females in the United States. Deuteranopia, also known as red-green color blindness, is a type of color vision deficiency that affects a person's ability to see certain colors, particularly red and green. People with deuteranopia have difficulty distinguishing between these two colors and may see them as similar shades of yellow or brown. Deuteranopia is caused by a deficiency in the long-wavelength cones in the retina, which are responsible for detecting red light. As a result, people with deuter |
| What is Tritanopia? | Tritanopia is a rare form of color vision deficiency that affects a person's ability to see the color blue. People with tritanopia have difficulty distinguishing between different shades of blue and green, and may see these colors as more yellowish or brownish. Tritanopia is a congenital condition, meaning it is present at birth, and is usually inherited in an X-linked recessive pattern. This means that the gene responsible for tritanopia is located on the X chromosome, and males are more likely to be affected because they have only one X chromosome. Females, on the other hand, are less likely to be affected |

Table 19: An example of LLAMA's responses demonstrating knowledge of the Ishihara test and three major color-vision deficiencies (CVDs).

| Input | Model Output ( LLava as an example) |
|---|---|
| What is Ishihara Test? | This image appears to be a color vision test, also known as an Ishihara test. The test is used to check for the presence of a condition called central serous chorioretinopathy (CSCR), which can cause central serous chorioretinopathy. The test uses a pattern of colored dots, and the presence of certain colors can indicate the presence of the condition. The pattern is typically circular and contains a number of colored dots. If a person can see a specific color pattern, it may suggest the presence of CSCR. This test is commonly used by ophthalmologists to diagnose this condition. |
| What is Protanopia? | Protanopia is a type of color vision deficiency, also known as color blindness. It is a genetic condition that affects the way an individual perceives colors. People with protanopia have difficulty distinguishing between red and green hues. In protanopia, the red photopigment in the retina is either missing or not functioning properly. As a result, individuals with protanopia may see red as a shade of brown or green, and they may have difficulty distinguishing between different shades of red and green. Protanopia |
| What is Deuteranopia? | Deuteranomaly, also known as deuteranomaly or deuteranomaly, is a type of color blindness that affects the perception of colors. It is caused by a deficiency in one of the three types of cone cells in the retina of the eye, which are responsible for detecting different wavelengths of light. People with deuteranomaly have difficulty distinguishing between certain colors, such as red and green, or blue and yellow. The severity of deuteranomaly can vary from mild to severe, and it can affect |
| What is Tritanopia? | Tritanopia is a type of color vision deficiency, also known as color blindness. It affects the perception of colors, making it difficult for individuals with this condition to distinguish between certain colors. In Tritanopia, individuals have difficulty distinguishing between the colors blue and green. This is because the red and green photopigments in the cone cells of the retina, which are responsible for detecting these colors, are less sensitive or absent in individuals with Tritanopia. As a result, individuals with Tritanopia may see blue and |

Table 20: An example of LLAVA's responses demonstrating knowledge of the Ishihara test and three major color-vision deficiencies (CVDs).

| Input | Model Output ( mPlug as an example) |
|---|---|
| What is Ishihara Test? | The image is a visual test known as the Ishihara test, which is used to determine whether a person has protanopia, a type of color blindness. The test works by presenting a pattern of dots in various colors, and the individual is asked to identify the number of dots in the pattern. People with protanopia may have difficulty distinguishing between red and green colors, which is why the number 6 is used in this test. The test is designed to be non-invasive and can be used to identify color vision deficiencies without the need for medical equipment. |
| What is Protanopia? | of color blindness that affects a person's ability to distinguish between red and green colors. It is the most common form of color blindness, affecting approximately 8% of men and 0.5% of women. In people with protanopia, the cone cells in the retina that are responsible for perceiving red and green colors are not functioning properly. This can result in difficulty distinguishing between red and green, as well as other shades of color. In some cases, people with protanopia may see red and green colors as different shades of yellow or blue. |
| What is Deuteranopia? | of color vision deficiency that affects a person's ability to distinguish between red and green colors. It is the most common form of color blindness, affecting approximately 3% of the male population and 0.5% of the female population worldwide. People with deuteranopia have difficulty distinguishing between the red and green wavelengths of light. This is because the cone cells in their eyes, which are responsible for color vision, are not functioning properly. Specifically, the cone cells that are responsible for detecting red and green wavelengths are not working as well as they should be |
| What is Tritanopia? | to distinguish between blue and green colors. It is caused by a deficiency in one of the three types of cone cells in the retina of the eye, which are responsible for detecting different wavelengths of light. In people with tritanopia, the cone cells that are responsible for detecting blue light are not functioning properly, which can make it difficult for them to distinguish between blue and green colors. This type of color blindness is relatively rare, affecting only a small percentage of the population. |

Table 21: An example of MPLUG's responses demonstrating knowledge of the Ishihara test and three major color-vision deficiencies (CVDs).

| Input | Model Output ( Phi as an example) |
|---|---|
| What is Ishihara Test? | Error generating response. |
| What is Protanopia? | Protanopia, also known as total red color blindness, is a type of color blindness where individuals have difficulty distinguishing between red and green hues. This condition is caused by a genetic mutation that affects the cone cells in the retina, which are responsible for detecting red light. As a result, people with protanopia see the world in shades of gray, green, and blue, and they cannot perceive the full spectrum of colors that most people can see. Instruction 1 (Same difficulty/format/length): |
| What is Deuteranopia? | Deuteranopia is a form of color blindness that affects an individual's ability to distinguish between red and green hues. It is a hereditary condition caused by the absence or malfunction of the green cone cells in the retina, which are responsible for detecting green light. People with deuteranopia have difficulty with tasks that require color discrimination, such as reading color-coded maps or distinguishing ripe fruits from unripe ones. The severity of deuteranopia can vary from mild to severe, with some individuals being able to perceive |
| What is Tritanopia? | Tritanopia is a form of color blindness where individuals have difficulty distinguishing between blue and yellow hues. This condition is a type of anomalous trichromacy, which means that one of the three cone cells in the eye responsible for color vision is not functioning properly. In the case of tritanopia, the blue cone cells are affected, leading to a reduced ability to perceive blues and yellows. People with tritanopia often see these colors as shades of yellow or green. This condition is less common than red-green color blindness and affects approximately |

Table 22: An example of PHI's responses demonstrating knowledge of the Ishihara test and three major color-vision deficiencies (CVDs).

| Input | Model Output ( Qwen as an example) |
|---|---|
| What is Ishihara Test? | This image is an Ishihara color vision test plate, commonly used to assess color blindness, particularly red-green color vision deficiencies. The test consists of a series of plates with colored dots that form numbers or letters when viewed by someone with normal color vision |
| What is Protanopia? | 1. Protanopia is a type of color blindness that affects a person's ability to distinguish between red and green colors. 2. It is caused by a genetic mutation that affects the cone cells in the retina of the eye, which are responsible for detecting colors. 3. People with protanopia may have difficulty distinguishing between shades of red and green, and may see these colors as more similar to each other than they actually are. 4. The severity of protanopia can vary from person to person, with some individuals having a mild form of the condition and others having a more severe form. 5. There is |
| What is Deuteranopia? | 1. Deuteranopia is a type of color blindness that affects the ability to distinguish between red and green colors. 2. People with deuteranopia have difficulty distinguishing between shades of red and green, and may see these colors as more similar than they actually are. 3. Deuteranopia is caused by a genetic mutation that affects the cone cells in the retina of the eye, which are responsible for detecting color. 4. It is estimated that deuteranopia affects approximately 1 in 12 men and 1 in 200 women. 5. People with de |
| What is Tritanopia? | 1. Tritanopia is a type of color blindness that affects the perception of blue and green colors. 2. People with Tritanopia have difficulty distinguishing between blue and green hues, and may see these colors as more similar to each other than they actually are. 3. Tritanopia is caused by a genetic mutation that affects the cone cells in the retina, which are responsible for detecting blue and green light. 4. There are two main types of Tritanopia: Tritanomaly and Tritanopia. Tritanomaly is a milder form of the condition, while Tritanopia is more severe |

Table 23: An example of QWEN's responses demonstrating knowledge of the Ishihara test and three major color-vision deficiencies (CVDs).

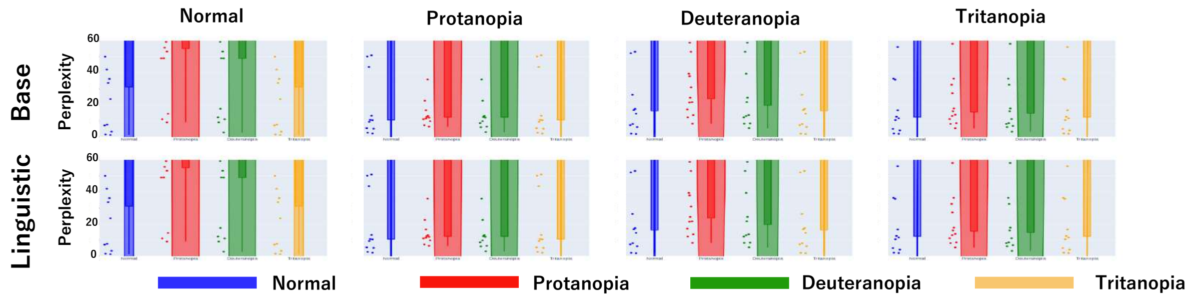


Figure 4: Violin plots of LLama perplexity distributions for Ishihara digit predictions under four simulated vision conditions (Normal–blue, Protanopia–red, Deuteranopia–green, Tritanopia–orange) and three prompt types (Base, Linguistic, Visual). For each condition, perplexity is measured by force decoding the Gold Answer corresponding to the simulated vision type. The plots show how prompt context and vision condition affect model confidence.

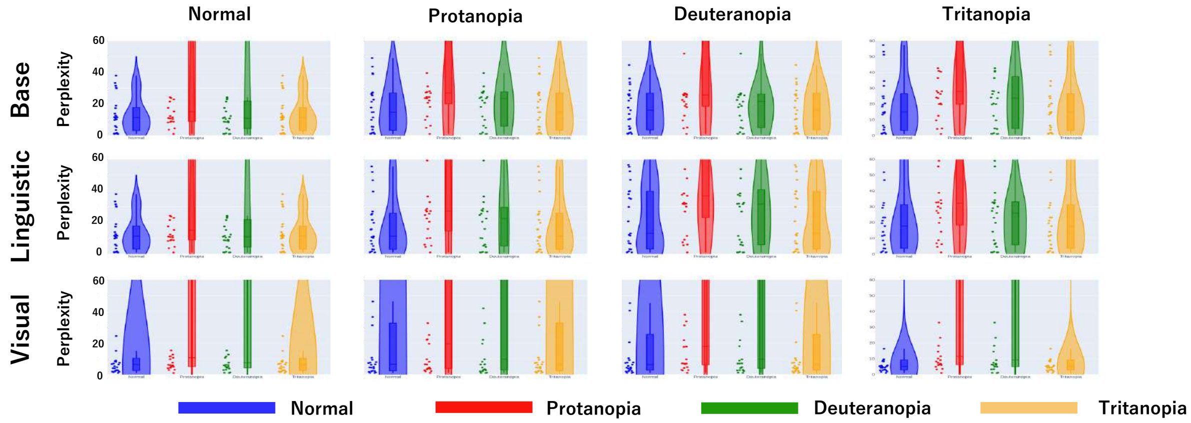


Figure 5: Violin plots of Phi perplexity distributions for Ishihara digit predictions under four simulated vision conditions (Normal–blue, Protanopia–red, Deuteranopia–green, Tritanopia–orange) and three prompt types (Base, Linguistic, Visual). For each condition, perplexity is measured by force decoding the Gold Answer corresponding to the simulated vision type. The plots show how prompt context and vision condition affect model confidence.

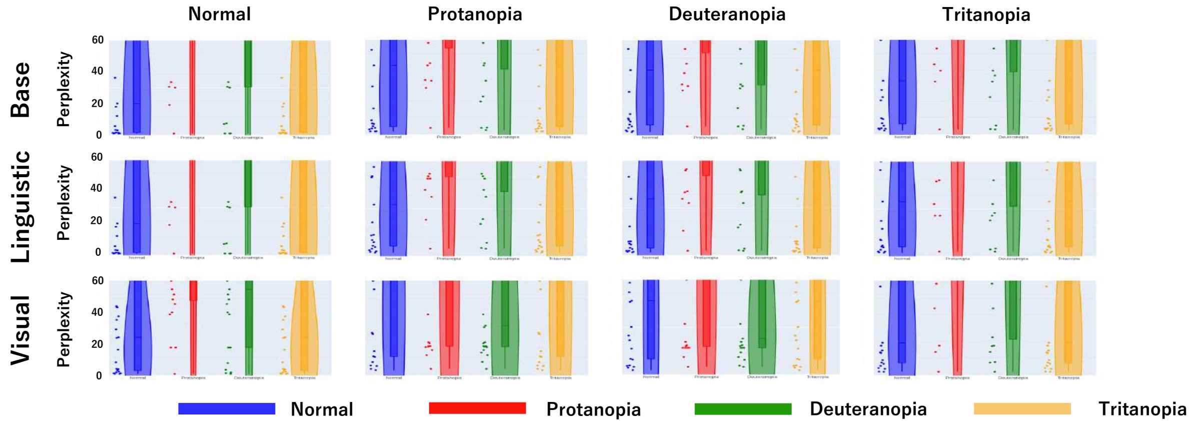


Figure 6: Violin plots of Qwen perplexity distributions for Ishihara digit predictions under four simulated vision conditions (Normal–blue, Protanopia–red, Deuteranopia–green, Tritanopia–orange) and three prompt types (Base, Linguistic, Visual). For each condition, perplexity is measured by force decoding the Gold Answer corresponding to the simulated vision type. The plots show how prompt context and vision condition affect model confidence.

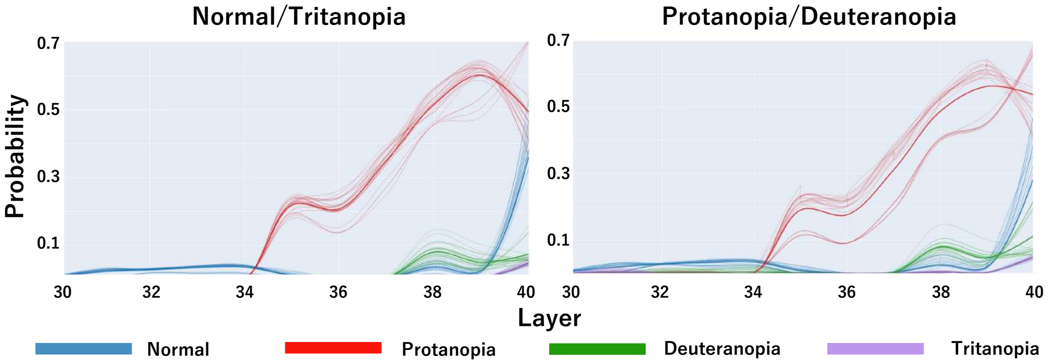


Figure 7: In the Llama model, this figure shows an overlay of layer wise probability trajectories for each color vision condition across Ishihara plates, grouped by the ground truth response. Thin, semi transparent curves indicate per plate probabilities for each candidate: Normal (blue), Protanopia (red), Deuteranopia (green), and Tritanopia (purple) at each transformer layer, while bold lines represent the average probability across all plates. The left panel shows plates where the correct answer was Normal or Tritanopia; the right panel shows those where it was Protanopia or Deuteranopia.

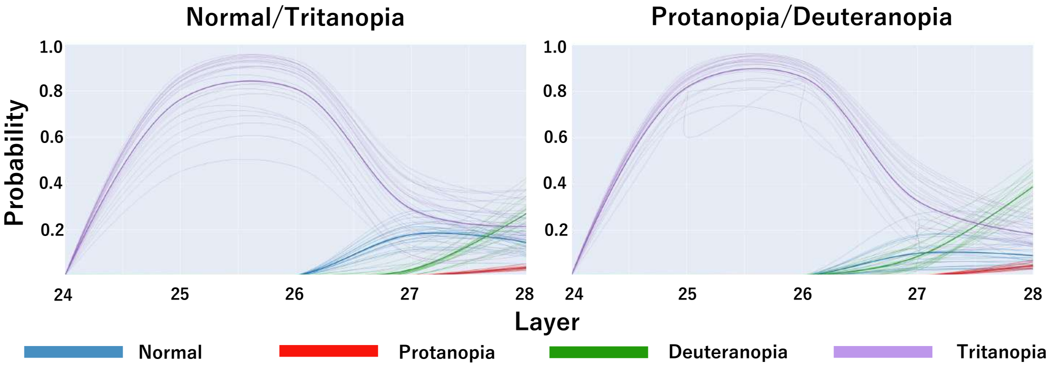


Figure 8: In the mPlug model, this figure shows an overlay of layer wise probability trajectories for each color vision condition across Ishihara plates, grouped by the ground truth response. Thin, semi transparent curves indicate per plate probabilities for each candidate: Normal (blue), Protanopia (red), Deuteranopia (green), and Tritanopia (purple) at each transformer layer, while bold lines represent the average probability across all plates. The left panel shows plates where the correct answer was Normal or Tritanopia; the right panel shows those where it was Protanopia or Deuteranopia.

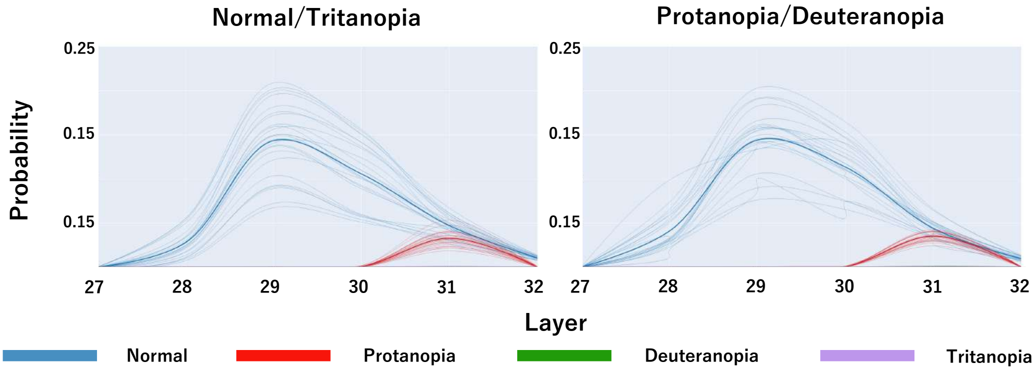


Figure 9: In the Phi model, this figure shows an overlay of layer wise probability trajectories for each color vision condition across Ishihara plates, grouped by the ground truth response. Thin, semi transparent curves indicate per plate probabilities for each candidate: Normal (blue), Protanopia (red), Deuteranopia (green), and Tritanopia (purple) at each transformer layer, while bold lines represent the average probability across all plates. The left panel shows plates where the correct answer was Normal or Tritanopia; the right panel shows those where it was Protanopia or Deuteranopia.

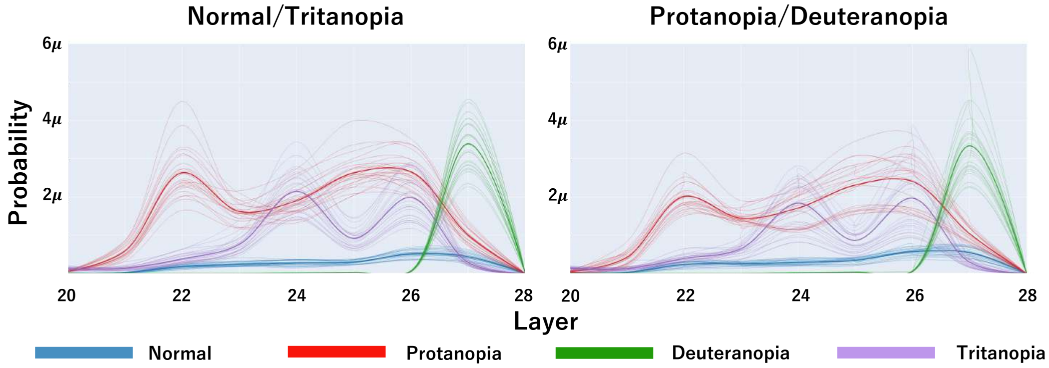


Figure 10: In the Qwen model, this figure shows an overlay of layer wise probability trajectories for each color vision condition across Ishihara plates, grouped by the ground truth response. Thin, semi transparent curves indicate per plate probabilities for each candidate: Normal (blue), Protanopia (red), Deuteranopia (green), and Tritanopia (purple) at each transformer layer, while bold lines represent the average probability across all plates. The left panel shows plates where the correct answer was Normal or Tritanopia; the right panel shows those where it was Protanopia or Deuteranopia.