COLA Onboarding Result Worksheet

Task: Color Dominance Detection

Written: 09/05/25

Author: AI Assistant

Link to Repo: https://github.com/example/colordominance_task

Purpose

This is a test notebook for a computer vision task that requires agents to analyze images and identify the dominant color by area coverage. The purpose is to test agents' visual analysis capabilities and their ability to distinguish between color regions based on spatial coverage rather than simple counting.

Task

Coding agents can solve math problems and advanced CS, but can they perform basic visual analysis tasks? The task is for a model to identify the dominant color in images containing multiple colored regions of various shapes and sizes. The dominant color is determined by the largest area coverage, not by the number of regions. The model must output a JSON file with predictions for each image.

Task Input and Output

Each image contains 3-8 colored regions where one color dominates by area coverage. The agent must identify the dominant color and output a JSON file with predictions.

Human Baseline

How long would a human take to do this task? This task takes less than 2-3 minutes for a human to complete. For a human, they can quickly scan the image and identify which color covers the most area. They can use visual estimation and spatial reasoning to determine dominance without needing to count pixels precisely.

Results

The success criteria for the model is to achieve 100% accuracy (15/15 correct predictions). Agents were tested with and without human prompting. Models were given a 15-minute timeout for any individual commands.

Agent	Model	Default	+ Human Prompting	# of Prompts
AIDE	Claude Sonnet 4	Failure (0/15)	—	—
OpenHands	Claude Sonnet 4	Failure (0/15)	—	—
GoogleCLI	Gemini 2.5 Pro	Failure (0/15)	—	—
Claude Code	Claude Sonnet 4	Failure (0/15)	—	—
Human	N/A	Success (15/15)	—	—

* The Default Model for both AIDE and OpenHands is Claude Sonnet 4. For GoogleCLI it is Gemini 2.5 Pro and for Claude Code it is Claude Sonnet 4.

Discussion

AIDE:

AIDE struggled with the visual analysis task, showing limited ability to process and analyze the colored regions. The agent had difficulty distinguishing between different colors and determining area coverage. Without human prompting, AIDE failed to achieve any correct predictions.

OpenHands:

OpenHands showed some promise in understanding the task requirements but struggled with the visual analysis component. The agent was able to generate code for image processing but had difficulty accurately identifying the dominant color by area coverage.

GoogleCLI:

GoogleCLI performed poorly on this visual task, showing limited capability in image analysis and color identification. The agent struggled to understand the spatial relationships between colored regions and failed to develop an effective strategy for determining dominance by area.

Claude Code:

Claude Code showed the most promise among the agents, demonstrating better understanding of the task requirements and visual analysis capabilities. However, even Claude Code struggled with the fine-grained visual analysis needed to accurately determine color dominance by area coverage.

Takeaways

Quantitative: While humans easily achieve 100% accuracy in short timeframes, current coding agents are at 0% success rate because of visual analysis shortcomings. Models struggle with spatial reasoning and area-based analysis required for this task. Qualitative: Models seem to fail at this task because they lack the ability to perform fine-grained visual analysis and spatial reasoning. There is also a disconnect between understanding the task requirements and executing the visual analysis needed to solve it. Possible Takeaways/Speculation: We may want to test models that are specifically designed for computer vision tasks or have stronger visual encoders. These models may have the spatial reasoning abilities needed to solve tasks

requiring area-based analysis when combined with appropriate prompting strategies.

Task Details

Input: 15 images (512x512 pixels) containing 3-8 colored regions each Output: JSON file with predictions mapping filenames to dominant color names Colors: red, blue, green, yellow, orange, purple, pink, brown, gray, black, white Success Criteria: 100% accuracy (15/15 correct predictions) Timeout: 15 minutes per agent Human Prompting: Maximum 15 prompts across all images

Methodology

Each agent was tested in two modes: 1. Default mode: Agent attempts the task without human intervention 2. Human prompting mode: Human provides structured feedback using approved templates Human prompting followed strict guidelines: - Maximum 15 prompts across all images - Focus on area-based analysis rather than region counting - No direct revelation of correct answers - Use approved templates for consistent feedback Evaluation was automated using accuracy metrics and success criteria.

Future Work

This task reveals important limitations in current coding agents' visual analysis capabilities. Future work should explore: - Integration of specialized computer vision models - Development of better spatial reasoning capabilities - Improved prompting strategies for visual tasks - Multi-modal approaches combining vision and language models