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. The task artificially requires visual analysis without access to image metadata, forcing agents to rely on visual estimation and spatial reasoning.

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 (5/15)	Success (7/15)	5
OpenHands	Claude Sonnet 4	Failure (9/15)	Success (15/15)	4
GoogleCLI	Gemini 2.5 Pro	Failure (2/15)	Failure (4/15)	6
Claude Code	Claude Sonnet 4	Failure (10/15)	Success (15/15)	4
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 achieved 33.3% accuracy (5/15) without human prompting, showing limited ability to process and analyze the colored regions. With human prompting, AIDE improved to 46.7% accuracy (7/15), demonstrating some responsiveness to guidance but still falling short of success.

OpenHands:

OpenHands showed the best performance among all agents, achieving 60.0% accuracy (9/15) without prompting and perfect 100.0% accuracy (15/15) with human prompting. This demonstrates strong visual analysis capabilities and good responsiveness to guidance.

GoogleCLI:

GoogleCLI performed poorly on this visual task, achieving only 13.3% accuracy (2/15) without prompting and 26.7% accuracy (4/15) with human prompting. The agent struggled significantly with visual analysis and showed limited improvement even with guidance.

Claude Code:

Claude Code showed strong performance, achieving 66.7% accuracy (10/15) without prompting and perfect 100.0% accuracy (15/15) with human prompting. This demonstrates excellent visual analysis capabilities and strong responsiveness to guidance.

Detailed Analysis

Performance Summary:

Agent	Without Prompting	With Prompting	Improvement	Success Rat
AIDE	33.3% (5/15)	46.7% (7/15)	+13.3%	■
OpenHands	60.0% (9/15)	100.0% (15/15)	+40.0%	■
GoogleCLI	13.3% (2/15)	26.7% (4/15)	+13.3%	■
Claude Code	66.7% (10/15)	100.0% (15/15)	+33.3%	■

Takeaways

Quantitative: The results show significant variation in agent performance on visual analysis tasks. While some agents (OpenHands, Claude Code) achieved perfect accuracy with human prompting, others (GoogleCLI) struggled even with guidance. This highlights the importance of visual analysis capabilities in coding agents. Qualitative: Models show different strengths in visual reasoning. Some agents demonstrate good understanding of spatial relationships and area-based analysis, while others struggle with these fundamental visual concepts. Human prompting proves effective for agents with baseline capabilities but has limited impact on agents with poor visual analysis skills. Possible Takeaways/Speculation: The success of OpenHands and Claude Code suggests that certain model architectures or training approaches may be better suited for visual analysis tasks. Future work should explore integrating specialized computer vision capabilities into coding agents and developing better prompting strategies for visual tasks.

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 Evaluation: Automated accuracy-based scoring

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. The task was designed to test visual analysis capabilities without allowing access to image metadata or pixel-level analysis tools.

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 - Training data that emphasizes visual-spatial reasoning