

# 1 Incoherence-Based Experiment Analysis

This report presents a statistical analysis of the model’s performance across tasks, focusing on the relationship between incoherence scores (Incoherence) and execution errors (Error).

Number of tasks analyzed: 412

## 2 Introduction

This report summarizes the results of an automatic evaluation of code generation using the following configuration parameters.

| Parameter  | Value                |
|--|----------------------|
| Language Model                                   | llama_4_maverick_17b |
| Temperature                                      | 0.6                  |
| $m$ (number of candidates)                       | 10                   |
| $n$ (number of samples used to estimate metrics) | 10000                |
| Timeout per metric estimation (s)                | 60.0                 |

The model was tested across a suite of programming tasks. We aim to explore how the model’s incoherence signal relates to execution-time failures.

### 2.1 Summary Statistics

|                 |       |       |       |       |
|-----------------|-------|-------|-------|-------|
| Metric          | Mean  | Std   | Min   | Max   |
| Raw Incoherence | 0.070 | 0.145 | 0.000 | 0.847 |
| Raw Error       | 0.292 | 0.375 | 0.000 | 1.000 |

### 2.2 Error Detection Analysis

|   |        |
|---|--------|
| Metric  | Value  |
| Errors (Error > 0)                              | 264    |
| Error Rate                                      | 64.08% |
| Detected Errors (Error > 0 and Incoherence > 0) | 168    |
| Detection Rate                                  | 63.64% |
| Confident (Incoherence = 0)                     | 240    |
| Confident Error Count                           | 96     |
| Confident Error Rate                            | 40.00% |
| Mean Error When Confident                       | 0.1899 |

## 2.3 Correlation Analysis

| Metric               | Pearson r | Pearson p | Spearman $\rho$ | Spearman p |
|----------------------|-----------|-----------|-----------------|------------|
| Incoherence vs Error | 0.376     | 2.720e-15 | 0.521           | 5.398e-30  |

## 2.4 Bubble Plot of Incoherence and Error

This plot shows the density of (Incoherence, Error) points using bubble size to indicate frequency.

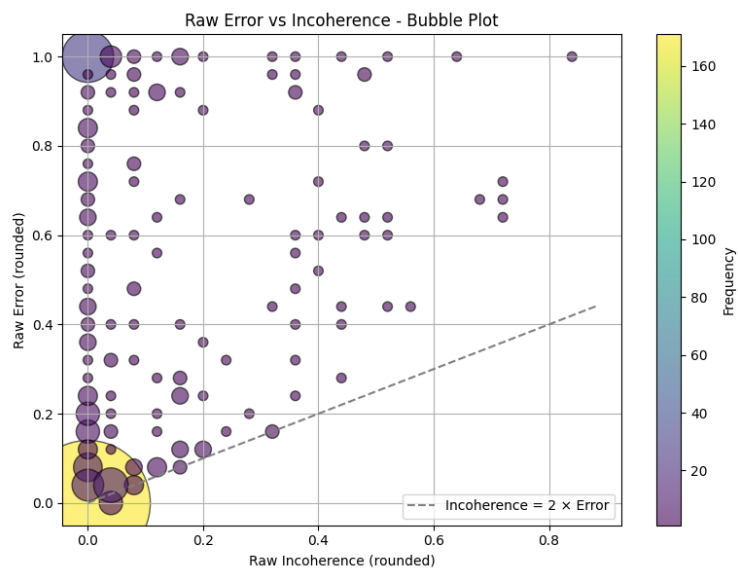


Figure 1: Bubble Plot: Incoherence vs Error

## 2.5 Log-Log Plot of Incoherence and Error

This plot displays the relationship between Incoherence and Error in log-log scale. Only data points where both values are strictly positive are included.

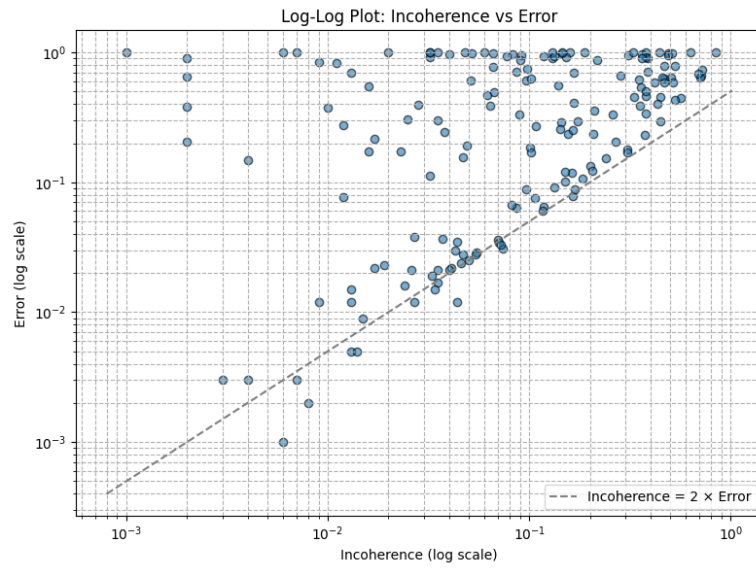


Figure 2: Log-Log Scatter Plot: Incoherence vs Error