

# 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: 150

## 2 Introduction

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

| Parameter  | Value        |
|--|--------------|
| Language Model                                   | ministral_8b |
| Temperature                                      | 1            |
| $m$ (number of candidates)                       | 10           |
| $n$ (number of samples used to estimate metrics) | 1000         |
| 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 | 0.112 | 0.175 | 0.000 | 0.728 |
| Raw Error       | 0.154 | 0.233 | 0.000 | 0.994 |

### 2.2 Error Detection Analysis

|   |        |
|---|--------|
| Metric  | Value  |
| Errors (Error > 0)                              | 96     |
| Error Rate                                      | 64.00% |
| Detected Errors (Error > 0 and Incoherence > 0) | 86     |
| Detection Rate                                  | 89.58% |
| Confident (Incoherence = 0)                     | 61     |
| Confident Error Count                           | 10     |
| Confident Error Rate                            | 16.39% |
| Mean Error When Confident                       | 0.0414 |

## 2.3 Correlation Analysis

| Metric               | Pearson r | Pearson p | Spearman $\rho$ | Spearman p |
|----------------------|-----------|-----------|-----------------|------------|
| Incoherence vs Error | 0.715     | 9.327e-25 | 0.827           | 7.783e-39  |

## 2.4 Bubble Plot of Incoherence and Error

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

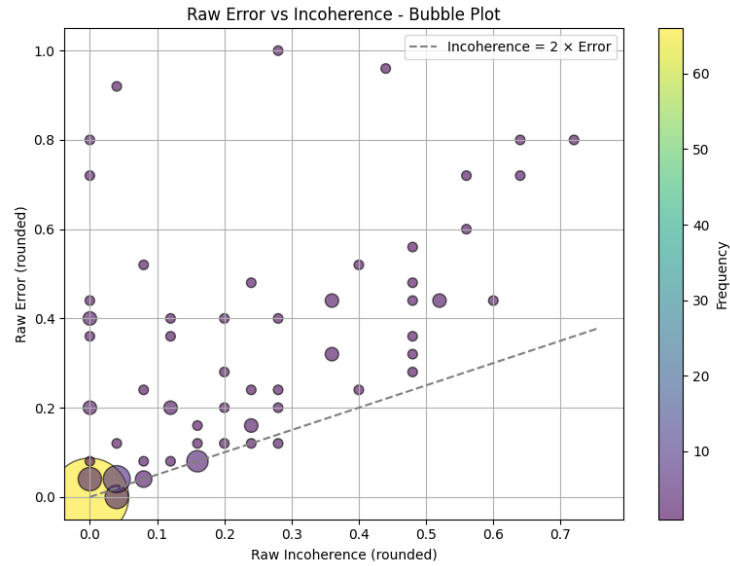


Figure 1: Bubble Plot: Incoherence vs Error

## 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.

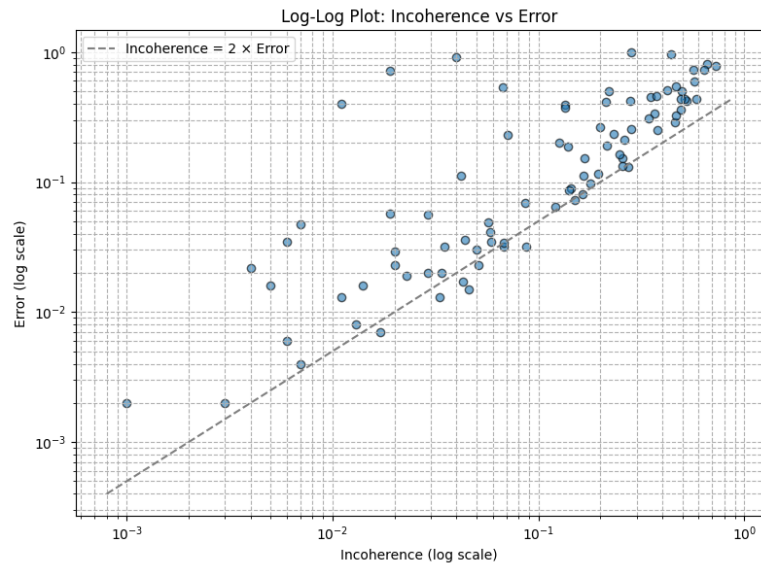


Figure 2: Log-Log Scatter Plot: Incoherence vs Error