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: 404

2 Introduction

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

Parameter	**Value**
Language Model	gpt_o4_mini
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	0.105	0.176	0.000	0.918
Raw Error	0.308	0.364	0.000	1.000

2.2 Error Detection Analysis

Metric	Value
Errors (Error $>$ 0)	285
Error Rate	70.54%
Detected Errors (Error $>$ 0 and Incoherence $>$ 0)	211
Detection Rate	74.04%
Confident (Incoherence = 0)	191
Confident Error Count	74
Confident Error Rate	38.74%
Mean Error When Confident	0.1800

2.3 Correlation Analysis

Metric	Pearson r	Pearson p	Spearman ρ	Spearman p
Incoherence vs Error	0.434	5.762e-20	0.567	8.623e-36

2.4 Bubble Plot of Incoherence and Error

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

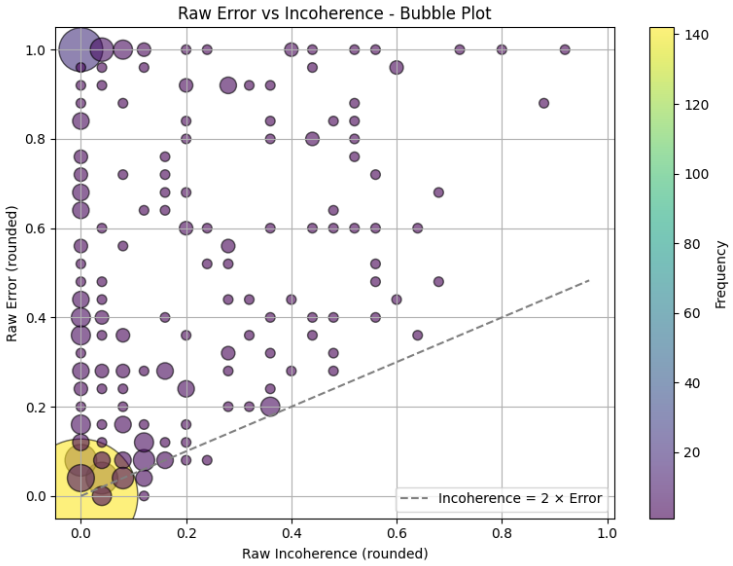

Figure 1: Bubble Plot: Incoherence vs Error

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