D Statistical Analyses for RQ1

This appendix provides the detailed results of the independent-samples t-tests conducted to evaluate the statistical significance of differences in error diversity, both between prompting strategies for each model (Table 4) and between the models themselves (Table 5). In both cases, we report Welch's t-statistic, associated degrees of freedom, p-values, and effect sizes using Hedges' g.

Prompting Comparisons within Models

Table 4 compares the three prompting strategies (IO, CoT, Self-Refine) within each model. Several key findings emerge:

- For **GPT-4o**, CoT prompts produced significantly more diverse erroneous code than either IO or Self-Refine (both $p < .001$, medium effect sizes). However, IO and Self-Refine did not differ significantly, suggesting CoT was the primary driver of diversity increases.
- For **GPT-5**, results were more mixed. CoT produced **significantly** higher diversity than IO ($\mathbf{p < .001}$, small effect), but did not differ significantly from Self-Refine. **IO and Self-Refine showed a significant difference, with IO producing greater diversity than Self-Refine ($\mathbf{p < .01}$).**
- For **Claude Sonnet 4**, **baseline** and **CoT** produced significantly higher diversity compared to **Self-Refine** (both $\mathbf{p < .001}$, medium effect sizes). Interestingly, **baseline** and **CoT** did not differ significantly ($\mathbf{p = 0.421}$).
- For **Gemini 2.5 Pro**, **two of the three** pairwise comparisons were statistically significant, though the effect sizes were generally small to medium. This indicates the model's diversity was sensitive to prompt framing, with CoT and Self-Refine producing more variation than IO. **However, the difference between Self-Refine and IO was not statistically significant ($\mathbf{p = 0.185}$).**
- For **Grok Code Fast 1**, CoT again yielded significantly higher diversity than both IO and Self-Refine, but IO and Self-Refine did not differ.

Taken together, these results highlight that the **influence of prompting strategy is model-dependent**: CoT consistently boosts error diversity for smaller models like GPT-4o and Grok Code Fast 1, while iterative refinement plays a larger role in Claude Sonnet 4. This suggests that the mechanisms by which LLMs generate variation are not uniform and may be linked to model architecture or training data.

D.1 Model Comparisons

Table 5 reports pairwise comparisons of mean edit distances across models, pooling over prompting strategies. The following trends emerge:

- **GPT-4o** consistently produced significantly less diverse erroneous code compared to all other models. Effect sizes were generally large negative (Hedges' g between -0.62 and -0.83) but **medium** against Grok Code Fast 1 ($\mathbf{g = -0.45}$). This confirms the descriptive finding that GPT-4o tends to generate conservative or repetitive error patterns.
- **Gemini 2.5 Pro** and **Claude Sonnet 4** exhibited the highest diversity overall. Their comparison with GPT-4o showed large and highly significant differences. **However, Claude Sonnet 4 was significantly less diverse than Gemini 2.5 Pro ($\mathbf{p < .001}$, $\mathbf{g = -0.11}$), indicating Gemini 2.5 Pro had the highest diversity.**
- **GPT-5** occupied an intermediate position, significantly more diverse than GPT-4o but **significantlylessdiverse** than Gemini 2.5 Pro ($\mathbf{p < .001}$, $\mathbf{g = -0.15}$). Interestingly, GPT-5 **didnotdiffersignificantly** from Claude Sonnet

4 (**p = 0.242**), but **wassignificantlymorediverse** than Grok Code Fast 1 (**p < .001**), indicating that it produced broader error distributions than Grok.

- **Grok Code Fast 1** tended to fall below Gemini and Claude in diversity, but above GPT-4o. Comparisons with Claude Sonnet 4 ($p < .001$, $g = \mathbf{0.26}$) and Gemini 2.5 Pro ($p < .001$, $g = 0.40$) indicate that Grok's diversity was significantly lower with **small** to moderate effect sizes (Group A is more diverse than Group B).

Overall, these statistical tests corroborate the descriptive results: **Gemini 2.5 Pro is the most diverse model, GPT-4o is the least diverse, and GPT-5, Claude Sonnet 4, and Grok Code Fast 1 fall in between**. Importantly, the effect sizes suggest these differences are not just statistically significant but also practically meaningful, especially in comparisons involving GPT-4o.

Table 4. Independent t-test results for prompt strategies by model.

Model	Group A	Group B	t-statistic	df	p-value	Hedges' g
GPT 4o	CoT	baseline	6.44	735.44	$2.17e-10$	0.43
GPT 4o	CoT	selfrefine	5.90	838.25	$5.38e-09$	0.37
GPT 4o	baseline	selfrefine	-0.40	1282.14	0.688	-0.02
GPT 5	CoT	baseline	3.70	2057.80	0.00023	0.16
GPT 5	baseline	selfrefine	-2.96	1119.27	0.0031	-0.15
GPT 5	CoT	selfrefine	0.12	1242.51	0.906	0.01
Claude Sonnet 4	baseline	selfrefine	6.43	1204.27	$1.79e-10$	0.33
Claude Sonnet 4	CoT	selfrefine	5.18	1176.87	$2.61e-07$	0.30
Claude Sonnet 4	CoT	baseline	-0.80	1359.72	0.421	-0.04
Gemini 2.5 Pro	CoT	selfrefine	4.16	750.48	$3.53e-05$	0.28
Gemini 2.5 Pro	CoT	baseline	2.81	810.28	0.0051	0.19
Gemini 2.5 Pro	selfrefine	baseline	1.33	1011.29	0.185	0.08
Grok Code Fast 1	CoT	baseline	4.08	575.84	$5.22e-05$	0.28
Grok Code Fast 1	CoT	selfrefine	3.23	650.64	0.0013	0.25
Grok Code Fast 1	baseline	selfrefine	-0.39	642.42	0.700	-0.03

Table 5. Independent t-test results for model comparisons.

Group A	Group B	t-statistic	df	p-value	Hedges' g
GPT 4o	Gemini 2.5 Pro	-22.97	2680.87	$1.01e-106$	-0.83
GPT 4o	GPT 5	-21.78	4396.18	$5.83e-100$	-0.62
GPT 4o	Claude Sonnet 4	-21.60	3930.99	$6.63e-98$	-0.67
GPT 4o	Grok Code Fast 1	-12.27	2730.24	$9.52e-34$	-0.45
Gemini 2.5 Pro	Grok Code Fast 1	10.62	2814.20	$7.27e-26$	0.40
Claude Sonnet 4	Grok Code Fast 1	8.00	3302.02	$1.66e-15$	0.26
GPT 5	Grok Code Fast 1	7.29	3220.87	$3.91e-13$	0.23
GPT 5	Gemini 2.5 Pro	-4.58	3073.81	$4.79e-06$	-0.15
Claude Sonnet 4	Gemini 2.5 Pro	-3.37	3223.77	0.00075	-0.11
GPT 5	Claude Sonnet 4	-1.17	4754.62	0.242	-0.03