E Statistical Analyses for RQ2

Prompting Comparisons within Models

As shown in Table 6, prompting strategy significantly affected alignment with authentic student errors, but the direction of the effect varied across models:

- For **GPT-4o**, Self-refine produced significantly worse alignment than either CoT or IO ($p < .001$, medium effects). This confirms the descriptive pattern that GPT-4o's alignment degrades under iterative refinement.
- For **GPT-5**, **Self-refine substantially reduced alignment** compared to both CoT ($\mathbf{p < .001}$, $\mathbf{g = 0.43}$) and IO ($\mathbf{p < .01}$, $\mathbf{g = 0.28}$), suggesting GPT-5's original outputs were **better** at simulating realistic mistakes.
- For **Claude Sonnet 4**, **CoT** and **IO** were **significantlybetter** than Self-refine ($\mathbf{p < .003}$ and $\mathbf{p < .02}$ respectively), though CoT and IO did not differ significantly ($\mathbf{p = 0.390}$). This reinforces Claude's overall robustness but indicates iterative refinement **decreases** alignment.
- For **Gemini 2.5 Pro**, results show that **CoT** ($\mathbf{p < .001}$, $\mathbf{g = 0.59}$) and $\mathbf{Self - refine}$ ($\mathbf{p < .001}$, $\mathbf{g = -0.49}$ when compared to IO) were **significantlybetter** than **IO** (baseline), with **CoTshowingthelargesteffectsize**.
- For **Grok Code Fast 1**, **Self-refine significantly reduced alignment** compared to both IO ($\mathbf{p < .001}$, $\mathbf{g = 0.60}$) and CoT ($\mathbf{p < .001}$, $\mathbf{g = 0.52}$). This suggests that $\mathbf{Self - refineledtotheworstmatch}$ to student errors.

Model Comparisons

Table 7 shows pairwise model comparisons:

- **GPT-5** consistently produced errors much **more similar** to students compared to all other models, with extremely large positive effect sizes (g between $\mathbf{0.35}$ and $\mathbf{1.79}$). This confirms GPT-5's position as the most consistent simulator of student-like mistakes.
- **Gemini 2.5 Pro** significantly outperformed Grok Code Fast 1 by a large margin ($\mathbf{g = 0.95}$) and was significantly **worse** than GPT-5 ($\mathbf{g = 0.35}$) and **ClaudeSonnet4** ($\mathbf{g = -1.29}$ when compared to Claude).
- **Claude Sonnet 4** significantly underperformed GPT-5 ($\mathbf{g = -1.79}$ when compared to GPT-5) and **Gemini2.5Pro** ($\mathbf{g = -1.29}$), but **outperformed** Grok Code Fast 1 ($\mathbf{g = -0.27}$ when compared to Claude) and GPT-4o ($\mathbf{g = 0.40}$).
- **GPT-4o** occupied a middle ground: **significantlyworse** than Claude ($\mathbf{g = 0.40}$) and **Gemini2.5Pro** ($\mathbf{g = -0.82}$), but **significantlybetter** than GPT-5 ($\mathbf{g = -1.26}$). Compared to Grok Code Fast 1, $\mathbf{GPT - 4o}$ showed a small, significant advantage ($\mathbf{g = 0.13}$).

In summary, these results statistically confirm that **GPT-5 provides the closest approximation of authentic student errors**, followed by a mixed group including Gemini 2.5 Pro, Claude Sonnet 4, GPT-4o, and Grok Code Fast 1. Prompting design modulates these effects, but model choice remains the dominant factor.

Table 6. Independent t-test results for prompt strategies by model (RQ2).

Model	Group A	Group B	t-statistic	df	p-value	Hedges' g
GPT 4o	CoT	selfrefine	-7.83	425.00	$3.85e-14$	-0.70
GPT 4o	CoT	baseline	-4.02	395.03	$6.92e-05$	-0.39
GPT 4o	baseline	selfrefine	-3.79	462.87	$1.68e-04$	-0.35
GPT 5	CoT	selfrefine	4.68	429.58	$3.79e-06$	0.43
GPT 5	baseline	selfrefine	3.09	433.91	$2.11e-03$	0.28
GPT 5	CoT	baseline	1.95	582.74	$5.14e-02$	0.16
Claude Sonnet 4	CoT	selfrefine	3.01	438.07	$2.77e-03$	0.28
Claude Sonnet 4	baseline	selfrefine	2.41	508.46	$1.63e-02$	0.21
Claude Sonnet 4	CoT	baseline	0.86	474.57	0.390	0.08
Gemini 2.5 Pro	CoT	baseline	5.93	390.17	$6.69e-09$	0.59
Gemini 2.5 Pro	baseline	selfrefine	-5.09	407.99	$5.36e-07$	-0.49
Gemini 2.5 Pro	CoT	selfrefine	1.15	395.34	0.249	0.11
Grok Code Fast 1	baseline	selfrefine	6.62	365.54	$1.26e-10$	0.60
Grok Code Fast 1	CoT	selfrefine	4.88	277.03	$1.78e-06$	0.52
Grok Code Fast 1	CoT	baseline	-1.41	396.83	0.158	-0.14

Table 7. Independent t-test results for model comparisons (RQ2).

Group A	Group B	t-statistic	df	p-value	Hedges' g
GPT 5	Claude Sonnet 4	36.08	1434.22	$2.13e-203$	1.79
GPT 5	Grok Code Fast 1	26.53	1343.54	$4.92e-125$	1.41
GPT 4o	GPT 5	-24.31	1449.66	$8.98e-110$	-1.26
Claude Sonnet 4	Gemini 2.5 Pro	-22.68	965.40	$1.20e-91$	-1.29
Gemini 2.5 Pro	Grok Code Fast 1	16.61	1161.37	$9.58e-56$	0.95
GPT 4o	Gemini 2.5 Pro	-14.65	1200.94	$7.45e-45$	-0.82
GPT 4o	Claude Sonnet 4	7.32	1187.92	$4.45e-13$	0.40
GPT 5	Gemini 2.5 Pro	6.53	1242.96	$9.66e-11$	0.35
Claude Sonnet 4	Grok Code Fast 1	-4.66	1064.99	$3.55e-06$	-0.27
GPT 4o	Grok Code Fast 1	2.25	1239.35	0.024	0.13