Model	Equivalent to developer	Code normalized equivalent to developer	Raises a conflict	Invalid output	Different resolution to developer
GPT 4.1	44.04%	54.09%	3.23%	0%	42.68%
Claude Opus 4	43.05%	49.38%	17.00%	8.06%	25.56%
Claude Sonnet 4	41.32%	48.26%	26.05%	0%	25.69%
Llama 4 Maverick	26.18%	32.63%	31.76%	.25%	35.36%
Llama 3.3 70B Instruct	1.86%	3.85%	81.02%	0%	15.13%
Gemini 2.5 Pro Preview	46.65%	53.35%	8.93%	.12%	37.60%
Qwen3 235B A22B	28.16%	35.73%	32.75%	.87%	30.65%
Grok 3 Beta	8.81%	11.66%	81.27%	0%	7.07%
QwQ 32B	24.07%	32.26%	13.77%	27.30%	26.67%
o3	49.63%	58.93%	3.10%	0%	37.97%
Qwen3 14B	12.90%	16.63%	69.48%	.12%	13.77%
Qwen3 32B	13.15%	16.87%	61.17%	.50%	21.46%
Deepseek R1 Distill Qwen 1.5B	0.00%	0.12%	0.00%	22.58%	77.30%
Deepseek R1 Distill Llama 8B	3.35%	7.57%	14.76%	5.83%	71.84%
Deepseek R1 Distill Qwen 14B	9.31%	13.40%	48.88%	.62%	37.10%
Deepseek R1 Distill Qwen 32B	22.83%	30.40%	30.65%	.99%	37.96%
Deepseek R1 Distill Llama 70B	25.81%	33.00%	29.40%	1.12%	36.48%
Deepseek R1	45.66%	53.60%	8.81%	.50%	37.09%
Ours	48.76%	58.93%	0.12%	0%	40.95%
Best SFT model	17.99%	23.70%	42.56%	1.74%	32.00%

Table 1: Merge-resolution performance across models. Top three results in each column are highlighted by color: 1st place, 2nd place, and 3rd place.