| Model | Correct merges | Semantic merges | Raising conflict | Valid Java markdown |
|---|---|---|---|---|
| GPT 4.1 | 44.04% | 54.09% | 3.23% | 100.00% |
| Claude Opus 4 | 43.05% | 49.38% | 17.00% | 91.94% |
| Claude Sonnet 4 | 41.32% | 48.26% | 26.05% | 100.00% |
| Llama 4 Maverick | 26.18% | 32.63% | 31.76% | 99.75% |
| Llama 3.3 70B Instruct | 1.86% | 3.85% | 81.02% | 100.00% |
| Gemini 2.5 Pro Preview | 46.65% | 53.35% | 8.93% | 99.88% |
| Qwen3 235B A22B | 28.16% | 35.73% | 32.75% | 99.13% |
| Grok 3 Beta | 8.81% | 11.66% | 81.27% | 100.00% |
| QwQ 32B | 24.07% | 32.26% | 13.77% | 72.70% |
| o3 | 49.63% | 58.93% | 3.10% | 100.00% |
| Qwen3 14B | 12.90% | 16.63% | 69.48% | 99.88% |
| Qwen3 32B | 13.15% | 16.87% | 61.17% | 99.50% |
| Deepseek R1 Distill Qwen 1.5B | 0.00% | 0.12% | 0.00% | 77.42% |
| Deepseek R1 Distill Llama 8B | 3.35% | 7.57% | 14.76% | 94.17% |
| Deepseek R1 Distill Qwen 14B | 9.31% | 13.40% | 48.88% | 99.38% |
| Deepseek R1 Distill Qwen 32B | 22.83% | 30.40% | 30.65% | 99.01% |
| Deepseek R1 Distill Llama 70B | 25.81% | 33.00% | 29.40% | 98.88% |
| Deepseek R1 | 45.66% | 53.60% | 8.81% | 99.50% |
| Ours | 48.76% | 58.93% | 0.12% | 100.00% |
| Best SFT model | 17.99% | 23.70% | 42.56% | 98.26% |

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