Compilation/Test Success Breakdown on DBCode–Bench
Default model settings

Outcome

- Failed to compile
- Compiled & tests failed
- Compiled & tests passed

Percentage of Generations

Model	Failed to compile	Compiled & tests failed	Compiled & tests passed
Claude Sonnet 4	10.9%	57.5%	31.6%
Gemini 2.5 Pro	14.8%	47.4%	37.8%
GPT–4o	34.8%	53.3%	11.9%
GPT–5	12.3%	47.9%	39.8%
gpt-oss–120b	39.0%	44.7%	16.3%
Kimi K2 0905	30.4%	48.4%	21.2%
o3	14.8%	53.6%	31.6%
Qwen3 Coder	17.3%	59.5%	23.2%