

Canonical Blind HumanEval Benchmark

164-problem blind execution benchmark evaluating Qwen2.5-Coder-14B via AtacamaODR port 8088.

Track 04: Canonical Blind HumanEval (164 Problems)

Evaluation Metric	Result	Reference / Baseline	Verification Standard
Pass@1 Accuracy	86.0% (141 / 164)	Claude 3.5: 93.7% GPT-4o: 90.2%	Subprocess Exit Code 0
Total Problems Evaluated	164 / 164 Problems	100% of OpenAI canonical suite	HumanEval.jsonl official
Logical / Assertion Failures	19 Problems	Subprocess assertion error	Captured traceback
Syntax / Parsing Failures	4 Problems	Subprocess SyntaxError	Captured traceback
Total Suite Execution Time	2,265.25s (~37.7 min)	Avg 13.8s per problem turn	Live wall-clock log

Audit Standard: Evaluated blindly against unmodified daemon port 8088. Zero eval-specific regexes. Every test assertion executed in an isolated subprocess.