

Canonical SWE-bench Lite Benchmark (300 Tasks)

300-task canonical execution benchmark evaluating Qwen2.5-Coder-14B on Apple Silicon using on-device Structural Context Compaction.

Track 05: SWE-bench Lite Fault Localization Ledger

Repository Target	Tasks	Accurately Localized	Localization Accuracy	Verification Standard
astropy/astropy	6	6	100.0%	Ground-truth diff match
django/django	114	105	92.1%	Ground-truth diff match
matplotlib/matplotlib	23	22	95.7%	Ground-truth diff match
mwaskom/seaborn	4	3	75.0%	Ground-truth diff match
pallets/flask	3	3	100.0%	Ground-truth diff match
psf/requests	6	3	50.0%	Ground-truth diff match
pydata/xarray	5	4	80.0%	Ground-truth diff match
pylint-dev/pylint	6	6	100.0%	Ground-truth diff match
pytest-dev/pytest	17	16	94.1%	Ground-truth diff match
scikit-learn/scikit-learn	23	22	95.7%	Ground-truth diff match
sphinx-doc/sphinx	16	14	87.5%	Ground-truth diff match
sympy/sympy	77	65	84.4%	Ground-truth diff match
Consolidated Total (12 Repos)	300 tasks	269 targets	89.7% Accuracy	100% On-Device (\$0.00)

Key Economic Finding: Average turnaround latency of 4.11s per repository task with \$0.00 cloud tokens. Acts as an on-device filter before escalating bounded repairs to frontier cloud models.