Multi-path coding hard screen

Cameron Sanderson · Exploratory · Local open-weight coding agents (Aider + Ollama; fair MLX)
n = 8 runs per path (mechanical, regression, two-file, scope × 2 seeds)

Headline

Most mid-size local models clear light coding cases and fail **multi-file coordination**. Only two paths finished perfect: a 14B coder and a 9B general model. One-shot harness repair did not rescue wrong model classes.

Path	Pass	Two-file	Wall	RSS
qwen2.5-coder:14b	1.00	2/2	~52s	~15 GB
qwen3.5:9b	1.00	2/2	~32s	~9–11 GB
qwen3-coder:30b	0.875	1/2	~20s	~18 GB
qwen3:14b / Gemma / mistral-small	0.75	0/2	varies	high
devstral	0.75	0/2 (timeouts)	~216s	~16 GB

Design rules

- One primary factor per phase (model or harness repair).
- External verify decides pass; agent narrative does not.
- Efficiency (wall + RSS) is a **second axis**, never averaged into quality.
- Sequential stacks only (no dual-load Ollama + MLX).

Repair (H2, one retry, two-file only)

Class	After H1	After one repair
Gemma / qwen3:14b	0/2	**0/2**
qwen3-coder:30b	1/2	**2/2**

Repair helps a near-miss coder. It does not convert the wrong model class into a multi-file default.

Defaults that followed the evidence

1. Multi-file reliability: `qwen2.5-coder:14b`
2. RAM tight + perfect suite: `qwen3.5:9b`
3. H2 only when the model is already near-miss

DEV / sealed-suite routes only. Not unsupervised production graduation.

Claim

Multi-file is the discriminating case on this instrument. Prefer model class over harness repair for multi-file failure. Keep pass rate and efficiency in separate columns.