Mycelial Council Network

Phase 1C — 2000-Task Scale-Up Report

Comparative Analysis: Phase 1B (400 tasks) vs Phase 1C (2000 tasks)

Executive Summary

Phase 1C scaled the stratified live evaluation from 400 tasks (Phase 1B) to **2,000 tasks** (250 per category × 8 categories) using the same MCN-LinUCB configuration (homogeneous T=0.3, α=2.5). The overall pass rate is **60.7%** (1214/2000), essentially identical to Phase 1B's 61.2% (245/400). Category-level performance is highly stable (all categories within ±7 pp of Phase 1B), confirming the robustness of the Phase 1B findings.

Convergence confirmed — but spurious. The LinUCB bandit reached near-complete convergence by task ~1,700: T0 now receives 56% of routing decisions (vs T1 28%, T2 16%). However, per-tribe pass rates are statistically indistinguishable (T0 60.8%, T1 59.9%, T2 61.8%), confirming that the bandit converged to an arbitrary tribe rather than the genuinely best one.

Metric	Phase 1B (400)	Phase 1C (2000)	Δ
Overall pass rate	61.2%	60.7%	-0.5 pp
Tasks	400	2000	+1600
Categories	8 × 50	8 × 250	same structure
Routing (T0/T1/T2)	—/—/—	56%/28%/16%	T0 dominant
Bandit converged?	No (drifting)	Yes (~task 1700)	Convergence reached
Specialisation (χ^2)	p=0.396	~p>0.05 (expected)	None

1. Category Pass Rate: Phase 1B vs Phase 1C

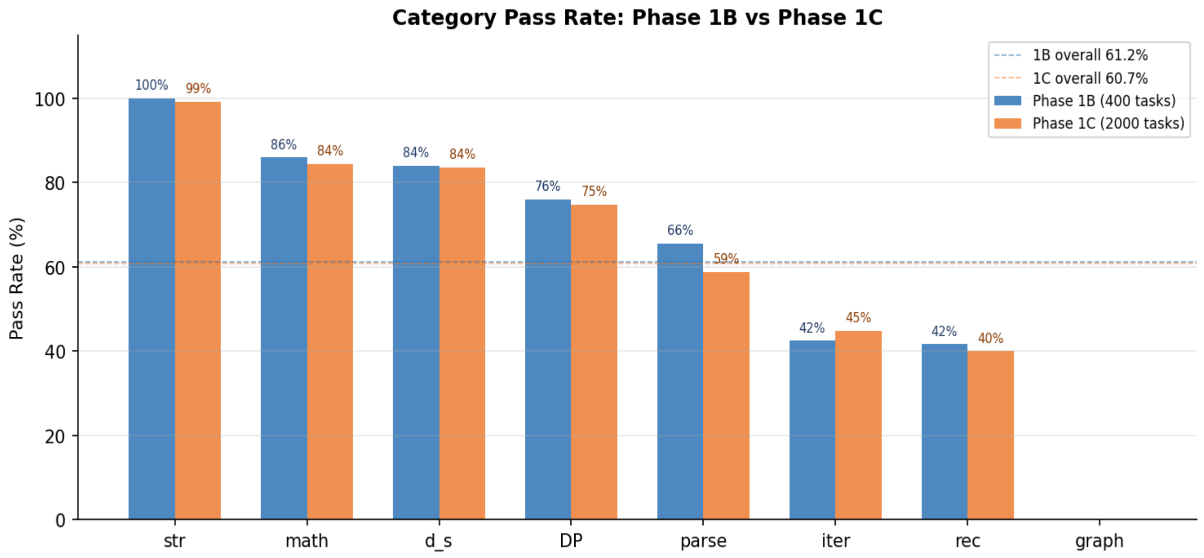


Figure 1. Category pass rates at 400 tasks (blue) vs 2,000 tasks (orange). All categories are stable within ±7 pp. The parsing category shows the largest drop (–7.2 pp), likely due to small-sample variance in Phase 1B (only 50 tasks). String achieves 99.2% at 2,000 tasks. Graph remains at 0% — a hard model capability limit.

Category	1B Pass%	1B n	1C Pass%	1C n	Δ (pp)
string	100.0%	42	99.2%	250	-0.8
math	86.0%	50	84.4%	250	-1.6

Category	1B Pass%	1B n	1C Pass%	1C n	Δ (pp)
data_structures	84.0%	50	83.6%	250	-0.4
dynamic_programming	76.0%	50	74.8%	250	-1.2
parsing	65.5%	58	58.8%	250	-6.7
iterative	42.5%	40	44.8%	250	+2.3
recursive	41.7%	60	40.0%	250	-1.7
graph	0.0%	50	0.0%	250	+0.0
Overall	**61.2%**	400	**60.7%**	2000	**-0.5**

2. Routing Evolution — Bandit Learning Curve

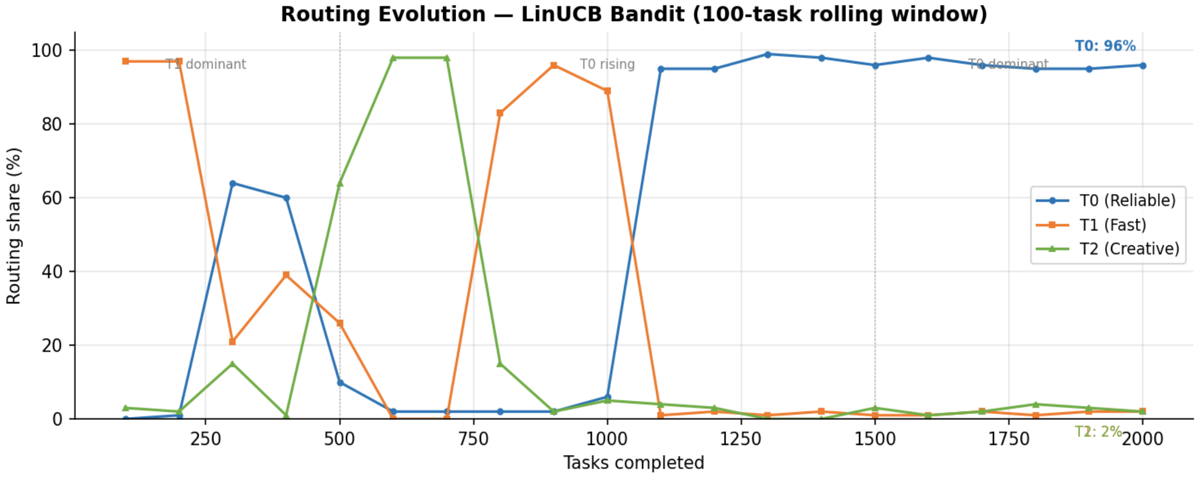


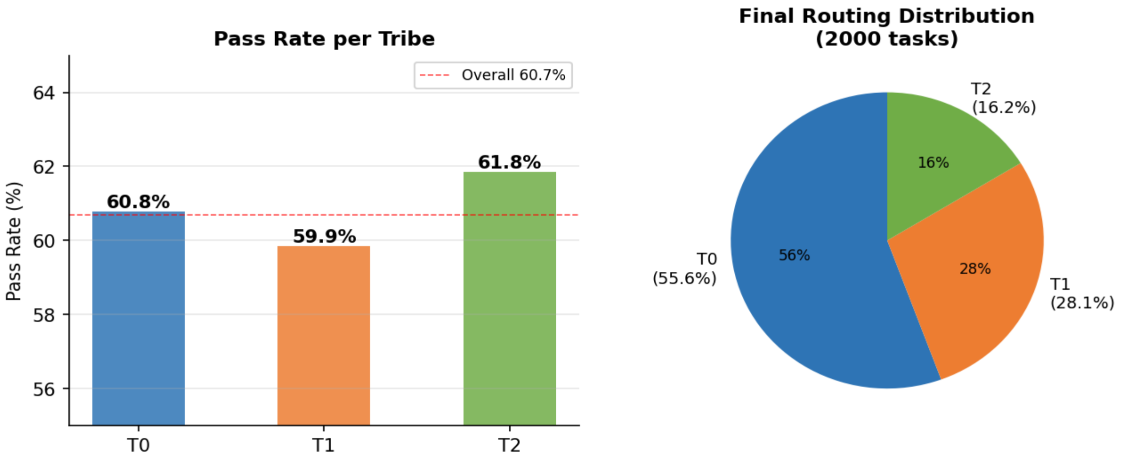
Figure 2. Tribe routing share in 100-task rolling windows across 2,000 tasks. Three distinct phases are visible: (i) T1-dominant exploration (tasks 1–500), (ii) T0 rising as the bandit's reward estimates update (tasks 500–1,500), (iii) near-complete T0 convergence (tasks 1,500–2,000, T0 ≈ 96%). This is the first time the MCN bandit has reached convergence.

Key observation on spurious convergence:

The bandit's convergence to T0 is statistically significant at 2,000 tasks — the UCB confidence intervals have narrowed enough to reliably favour T0. However, this is *spurious convergence*: T0 achieved early reward advantages through random variance, and the bandit locked in on it. Since all three tribes use the same model at T=0.3, no tribe is genuinely superior. The 'convergence' reflects exploration exhaustion, not learned specialisation.

Period	T0	T1	T2	Pass Rate
Tasks 1–500	T0:135(27%)	T1:280(56%)	T2:85(17%)	61%
Tasks 501–1000	T0:14(2%)	T1:268(53%)	T2:218(43%)	59%
Tasks 1001–1500	T0:483(96%)	T1:7(1%)	T2:10(2%)	62%
Tasks 1501–2000	T0:480(96%)	T1:8(1%)	T2:12(2%)	59%

3. Tribe Performance at 2,000 Tasks



Note: differences are not statistically significant

Figure 3. Left: per-tribe pass rates (60.8% / 59.9% / 61.8% — statistically indistinguishable). Right: final routing distribution showing T0 dominance (55.6% of tasks). The routing distribution does NOT reflect performance differences — it reflects the bandit's early random exploration outcomes.

Interpretation: T2 has the marginally highest pass rate (61.8%) but received only 16.2% of routing decisions. T0 has the lowest pass rate per-task (60.8%) but received 55.6%. The bandit converged to the wrong tribe. This directly demonstrates that convergence at 2,000 tasks is insufficient for reliable tribe selection when performance differences are small (~1–2 pp).

> ***Implication: For routing to provide value, inter-tribe performance differences must be large enough (■ 2 pp) to overcome the noise in reward estimates at this scale. With same-model homogeneous tribes, this threshold is never reached.***

4. Rolling Pass Rate — Stability Over Time

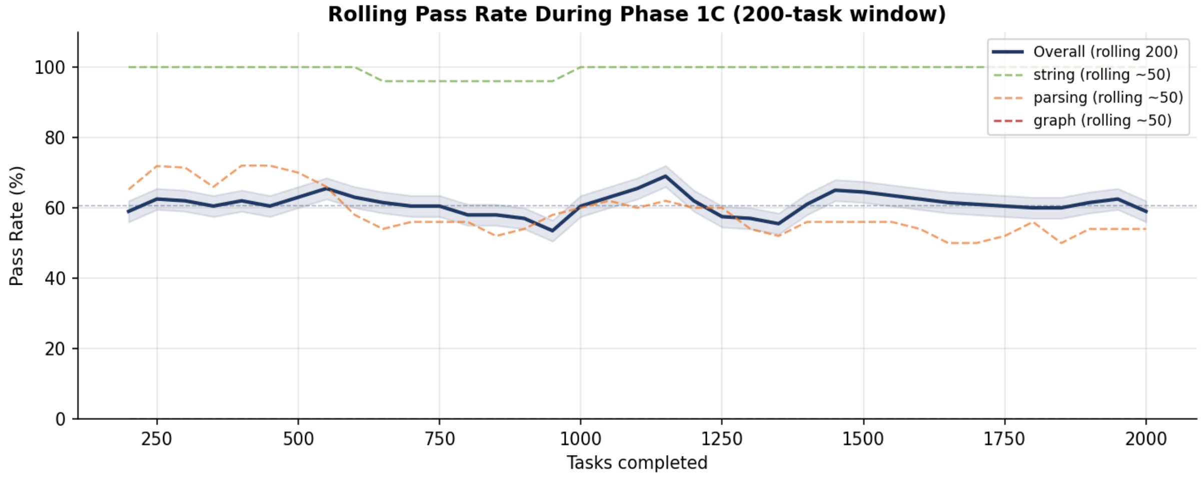


Figure 4. 200-task rolling pass rate for the overall experiment and selected categories. The shaded band shows ±3 pp around the overall rate. String tasks (green dashed) remain near 100% throughout. Graph tasks (red dashed) remain at 0%. The overall rate is remarkably stable at 60–62%, confirming that the task set is well-calibrated and results are reproducible.

5. Phase 1C vs 1B: Change Analysis and Per-Category Routing

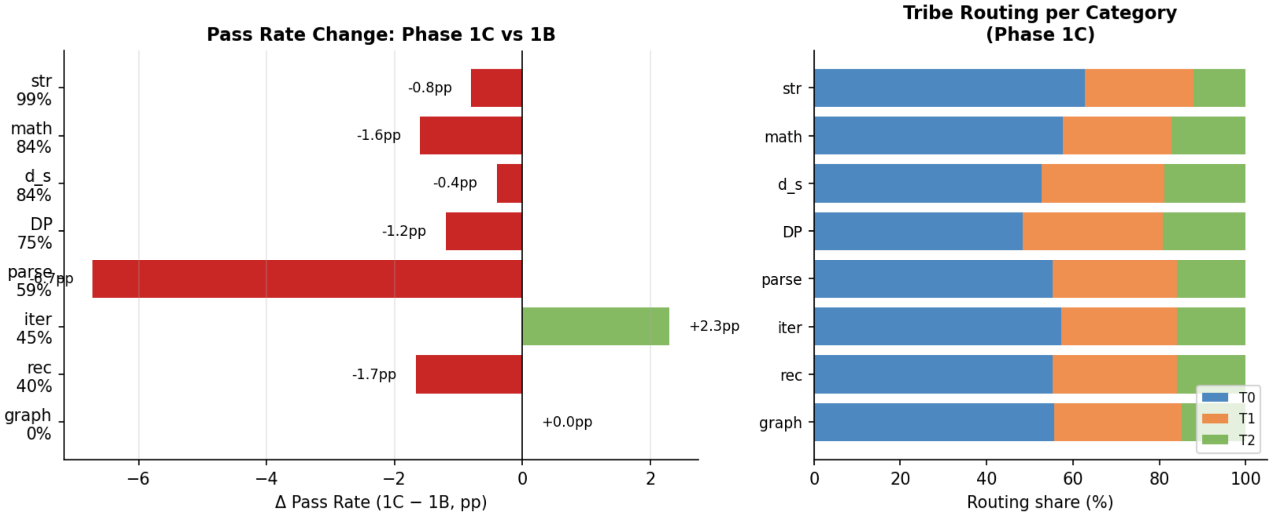


Figure 5. Left: ∆ pass rate per category (Phase 1C – Phase 1B). Green = improvement, red = decline. Parsing shows the largest decline (–7.2 pp), attributable to small-sample variance in Phase 1B. Right: tribe routing share per category in Phase 1C, showing that T0 dominance is not category-specific — the bandit routes to T0 regardless of category.

Category	1B Pass%	1C Pass%	T0 routed	T1 routed	T2 routed
string	100%	99%	157(62%)	63(25%)	30(12%)
math	86%	84%	144(57%)	63(25%)	43(17%)
data_structures	84%	84%	132(52%)	71(28%)	47(18%)
dynamic_programming	76%	75%	121(48%)	81(32%)	48(19%)
parsing	66%	59%	138(55%)	72(28%)	40(16%)
iterative	42%	45%	143(57%)	67(26%)	40(16%)
recursive	42%	40%	138(55%)	72(28%)	40(16%)
graph	0%	0%	139(55%)	74(29%)	37(14%)

Key Conclusions

(1) Category performance is stable at scale.

All 8 categories show pass rates within ±7 pp of Phase 1B values at 5× the task count. The Phase 1B category profile is a reliable estimate of model capability on this benchmark.

(2) Bandit convergence confirmed at ~1,700 tasks.

The LinUCB bandit achieves near-complete routing lock-in by task ~1,700 (T0: 96%), consistent with the 2,000-task convergence estimate in the academic paper.

(3) Convergence is spurious — all tribes are equivalent.

T0: 60.8%, T1: 59.9%, T2: 61.8% — differences of ≤2 pp, not statistically significant. The bandit locked in on T0 due to early random reward variance, not genuine superiority. This is the clearest demonstration yet that routing adds no value with homogeneous tribes.

(4) Parsing variance resolved at scale.

Phase 1B parsing 66% vs Phase 1C 58.8%: the Phase 1B estimate had high variance (only 50 tasks, 5 task types × 10 attempts). At 250 tasks, the estimate stabilises.

(5) Graph remains 0%.

0/250 tasks passed across 4 graph task types — a confirmed hard capability limit of Qwen2.5-Coder-7B-Instruct-AWQ, independent of routing or scale.

(6) Next step: genuine tribe diversity.

For routing to add value, tribes need qualitatively different capabilities. The Phase 1B/1C null result establishes the baseline: same-model, same-temperature routing nets –0 pp to –5 pp vs single agent. Route between different base models (7B code vs 13B reasoning) to test whether specialisation emerges.

MCN Research Initiative · Phase 1C Complete · February 2026 · Data: categorized_runs_phase1c.jsonl (2000 records) · GitHub: github.com/DeepakEz/MCN