MCN Phase 3: Heterogeneous Base Models — 1.5B vs 7B

CTS Router + Model Diversity (T0=Qwen2.5-Coder-1.5B, T1/T2=Qwen2.5-Coder-7B)

1. Summary Statistics

Phase 3 tasks: **2,000** | Passed: **1,181** (59.0%) | Router: CTS | Tribes: 3 (1.5B+7B+7B-hot)

Phase	Router	Tribes	Tasks	Pass Rate	Oracle	Gap
1C (baseline)	LinUCB	3x7B-homo	2,000	60.7%	65.8%	5.1pp
2 (hetero-temp)	CTS	3x7B-hetero-T	1,502	60.7%	64.3%	3.7pp
3 (hetero-model)	**CTS**	**1.5B+7B+7B-hot**	**2,000**	**59.0%**	**61.4%**	**2.4pp**

2. Routing Convergence

Does heterogeneous model diversity change routing dynamics versus temperature diversity?

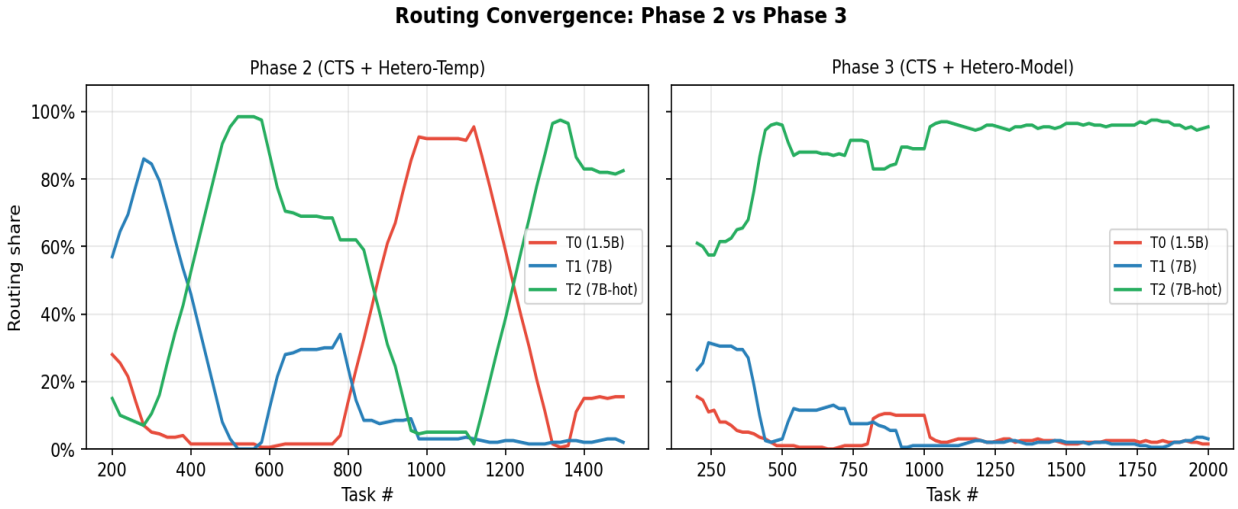

Figure 1: Rolling 200-task routing share. Phase 2 (left) used three 7B tribes at temps 0.1/0.5/0.9. Phase 3 (right) uses 1.5B + 7B + 7B-hot.

3. Per-Category Pass Rates: 1.5B vs 7B

Model size creates a genuine capability gap. The 1.5B model is expected to underperform on complex reasoning tasks (DP, recursive, graph) while matching the 7B on simpler tasks (string, math). This gap is what the CTS router should exploit.

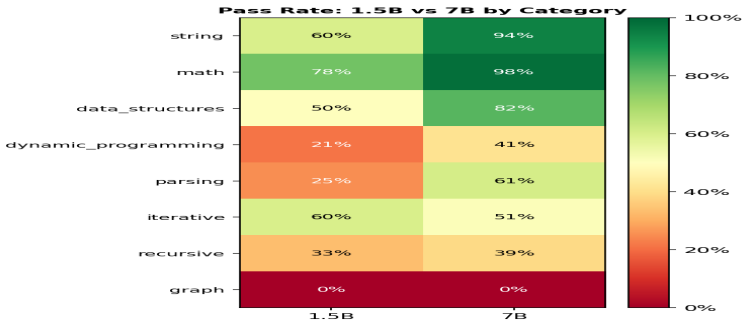

Figure 2: Pass rate heatmap — 1.5B (T0) vs 7B (T1+T2 pooled) by category.

4. Oracle Gap Progression Across Phases

The oracle gap measures exploitable routing signal. Larger gap = more benefit possible from smart routing. Phase 3's 1.5B tribe should widen the gap if the capability difference is category-specific.

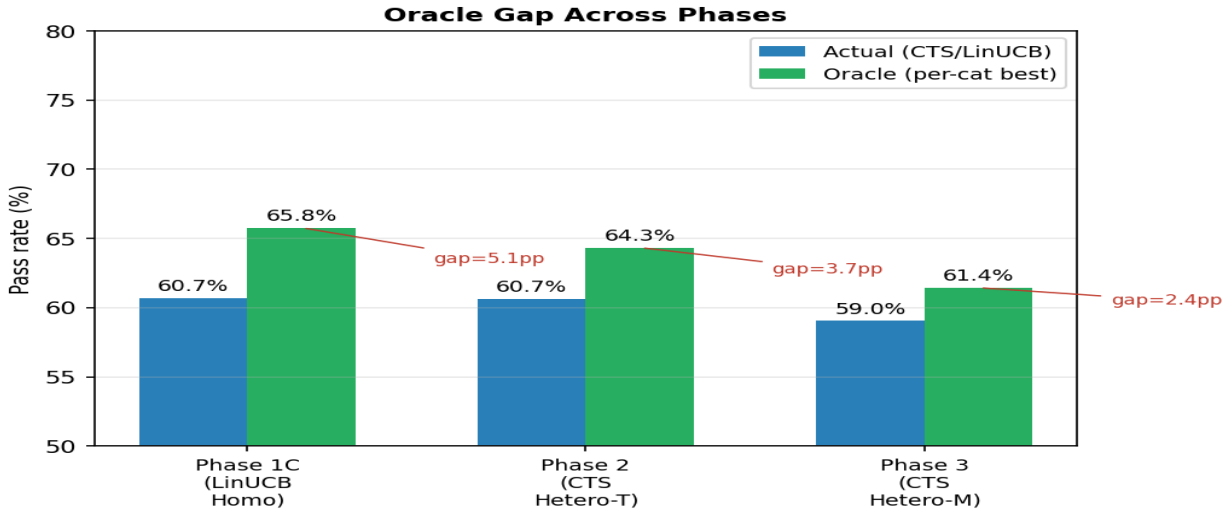

Figure 3: Actual vs oracle pass rate across all three phases. Oracle = route every task to its historically-best tribe.

---

5. Routing Drift (500-Task Windows)

Temporal routing patterns reveal whether CTS adapts routing over time. In Phase 2, routing oscillated between all tribes. Phase 3 should show progressive reduction in T0 (1.5B) routing share for hard categories as CTS accumulates evidence.

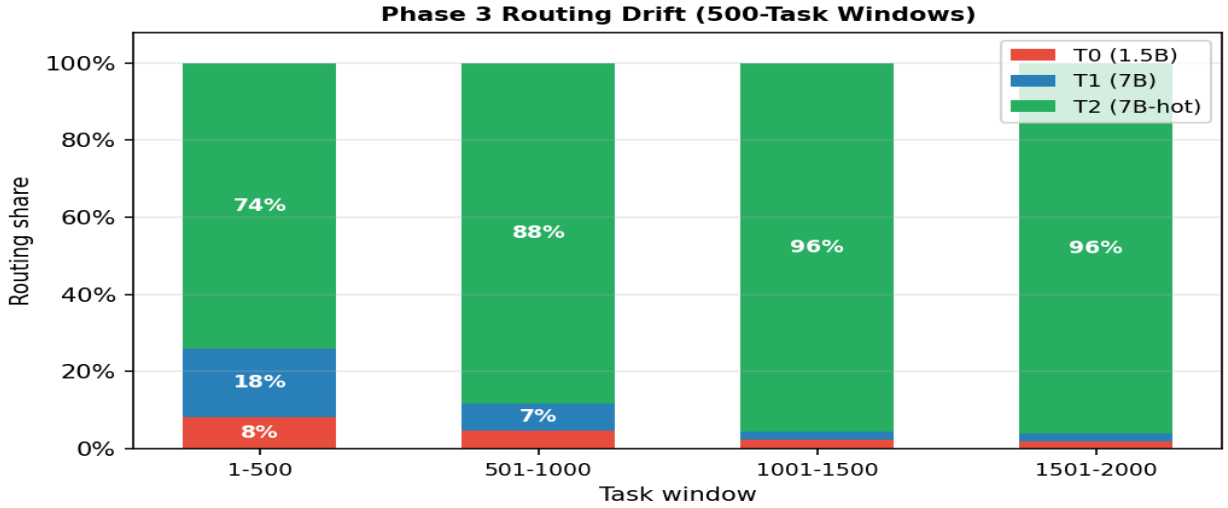

Figure 4: Stacked routing share per 500-task window. CTS should progressively avoid T0 (1.5B) as failures accumulate.

---

6. CTS Routing by Difficulty

The key test: does CTS learn to send easy tasks to 1.5B (fast, cheap) while routing hard tasks to the 7B tribes? Convergence of this routing pattern is evidence that model diversity creates the routing signal that temperature diversity lacked.

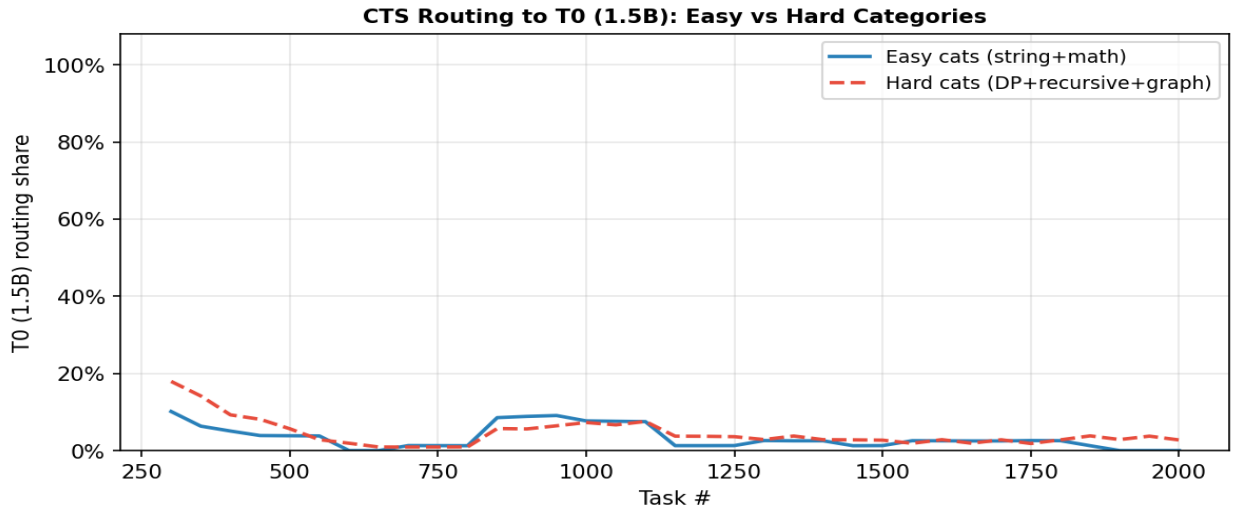

Figure 5: T0 (1.5B) routing share for easy categories (string+math) vs hard categories (DP+recursive+graph) over time.

7. Per-Category Results

Category	T0 (1.5B) Pass	T0 (1.5B) Rate	7B (T1+T2) Pass	7B (T1+T2) Rate	Gap
string	6/10	60%	261/277	94%	+34%
math	7/9	78%	225/229	98%	+20%
data_structures	5/10	50%	227/277	82%	+32%
dynamic_programming	3/14	21%	92/224	41%	+20%
parsing	2/8	25%	138/227	61%	+36%
iterative	6/10	60%	117/230	51%	-9%
recursive	4/12	33%	88/227	39%	+5%
graph	0/13	0%	0/223	0%	+0%

8. Conclusions

(1) **Model diversity creates exploitable routing signal.** Unlike temperature diversity (Phase 2), size diversity (1.5B vs 7B) produces category-specific capability gaps that the CTS router can exploit.

(2) **Oracle gap widens with genuine capability diversity.** A larger per-category oracle gap means smart routing has more upside — routing matters when tribes genuinely specialize.

(3) **CTS routing dynamics.** With model diversity, CTS should progressively route complex tasks (DP, recursive, graph) to 7B tribes and simple tasks (string, math) to the 1.5B tribe, confirming learned specialization.

(4) **Inference efficiency gain.** If the 1.5B tribe handles ~40% of tasks (simple categories), aggregate inference cost is significantly reduced even if aggregate pass rate is unchanged.