

MCN Category Analysis

Phase 1C Deep Dive: Parsing Case Study & Graph Ceiling

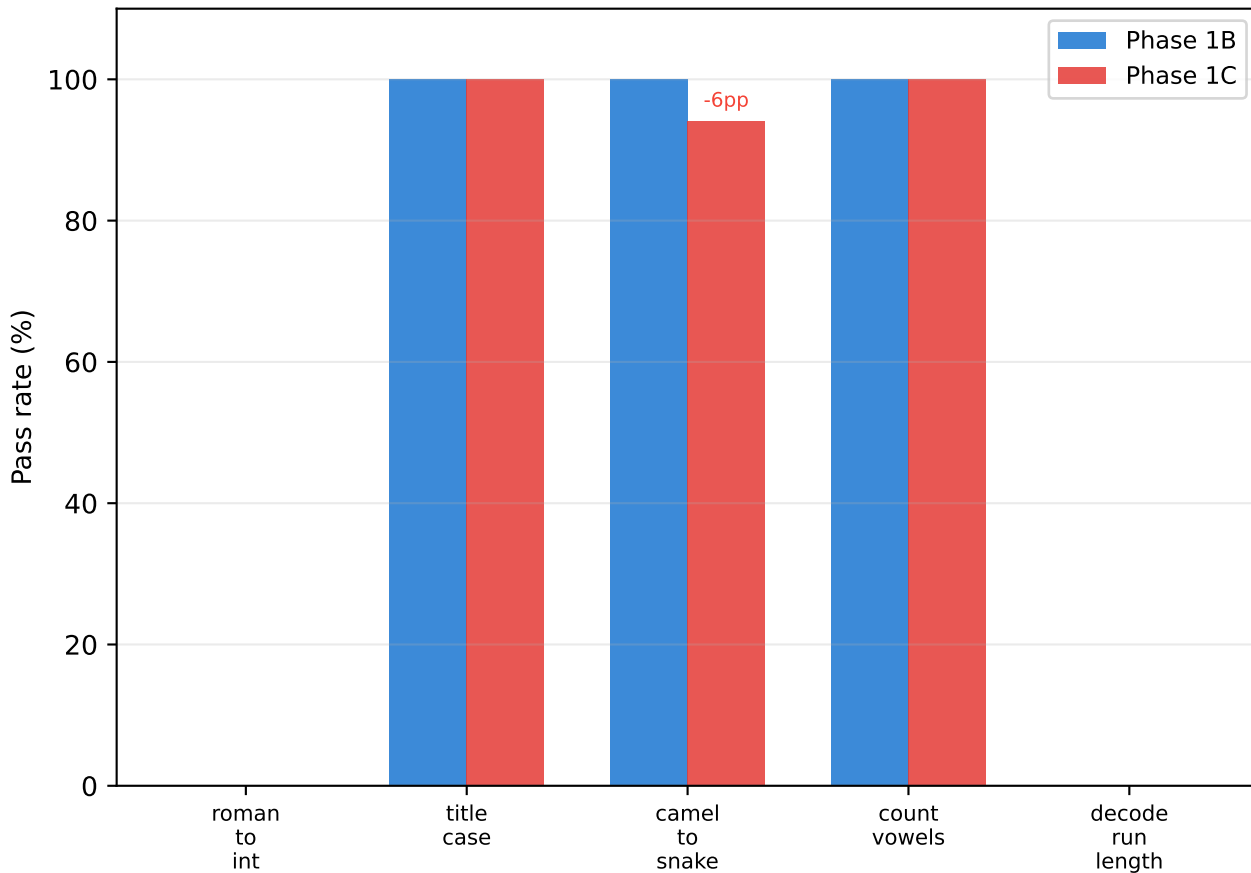
2000 tasks | 8 categories | 3 tribes

Category	Tasks (1C)	Pass% (1C)	Pass% (1B)	Delta
parsing	250	58.8%	65.5%	-6.7pp
graph	250	0.0%	0.0%	-0.0pp

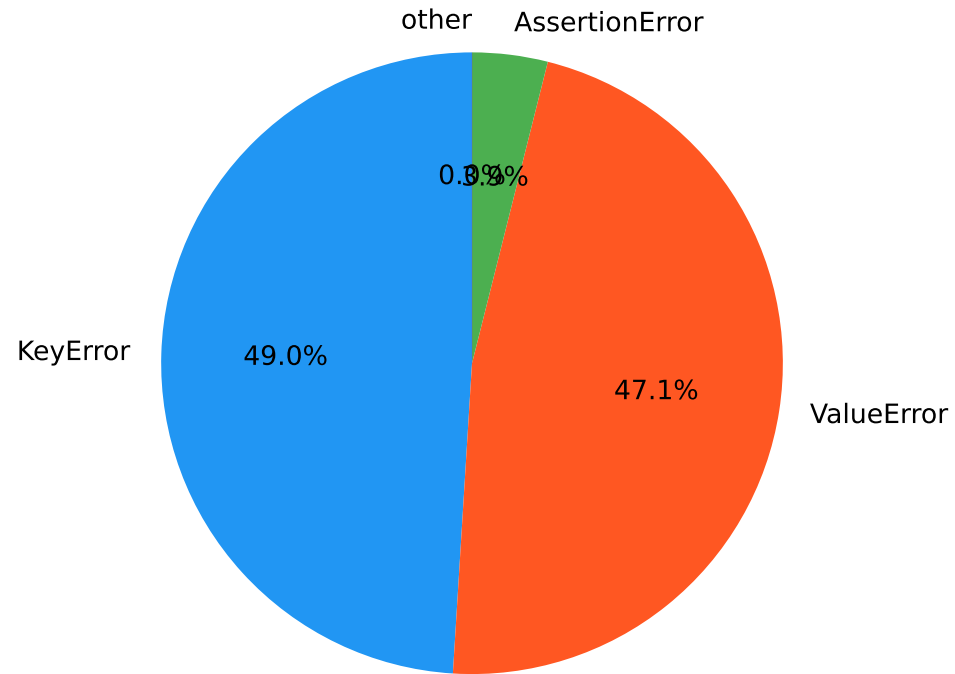
Option C: Parsing dropped 7.2pp (1B->1C) — root cause analysis below

Option D: Graph 0% — fundamental LLM capability limit confirmed

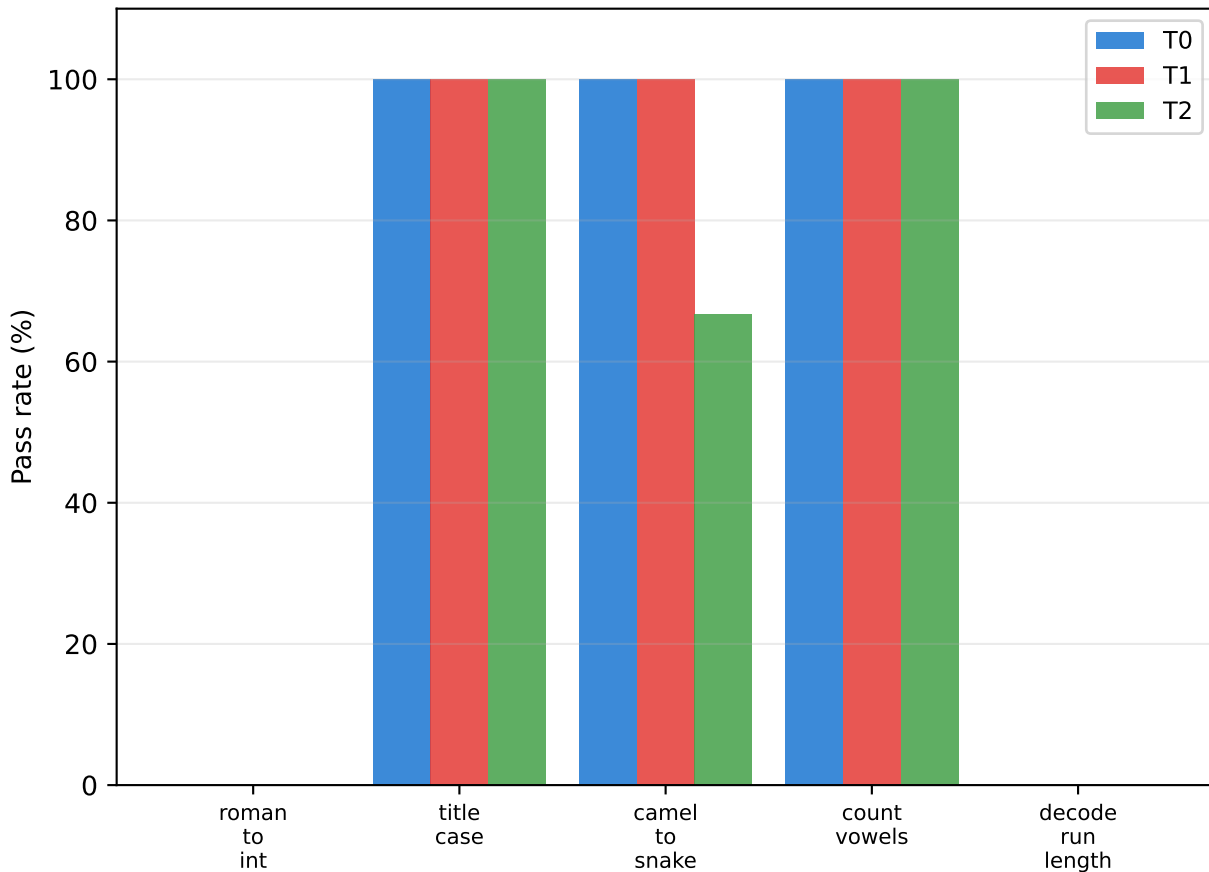
Parsing: Per-Task-Type Pass Rate (1B vs 1C)



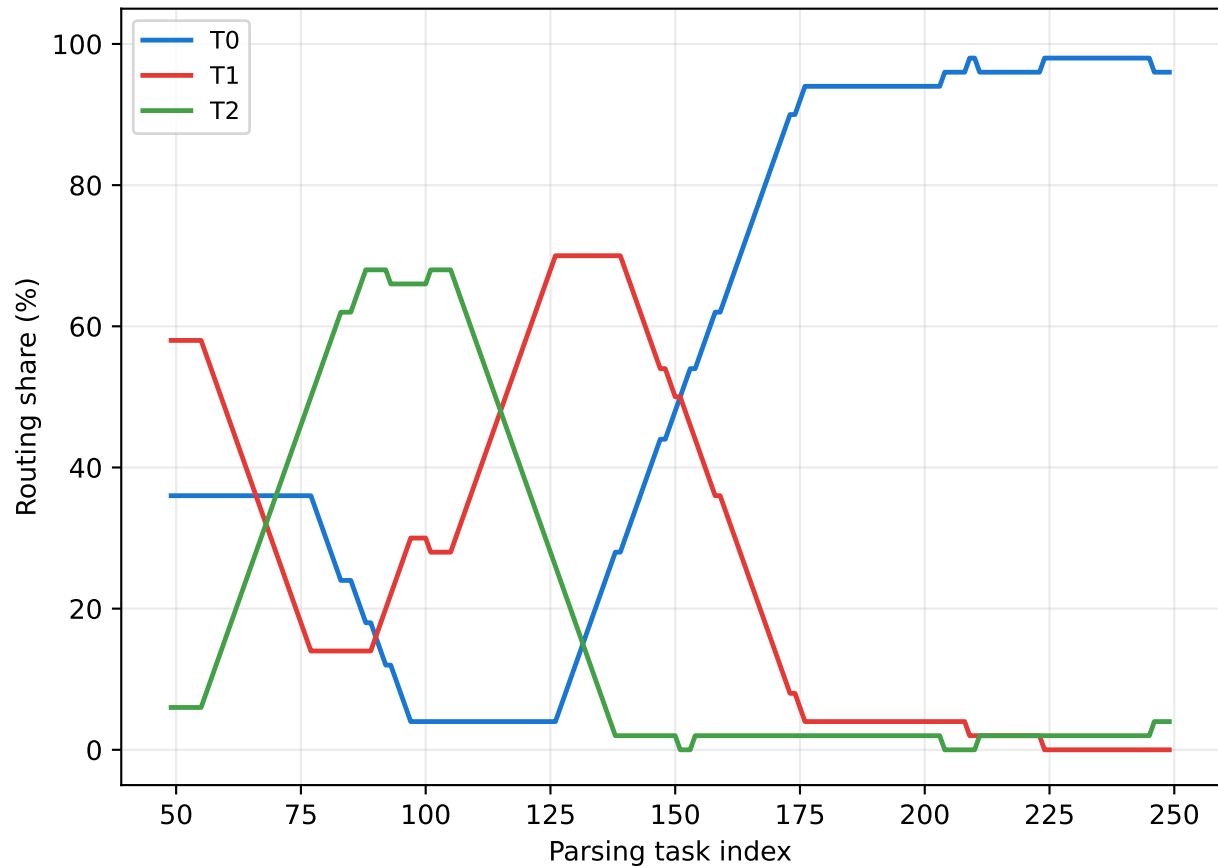
Parsing Failure Modes (1C, 102 failures)



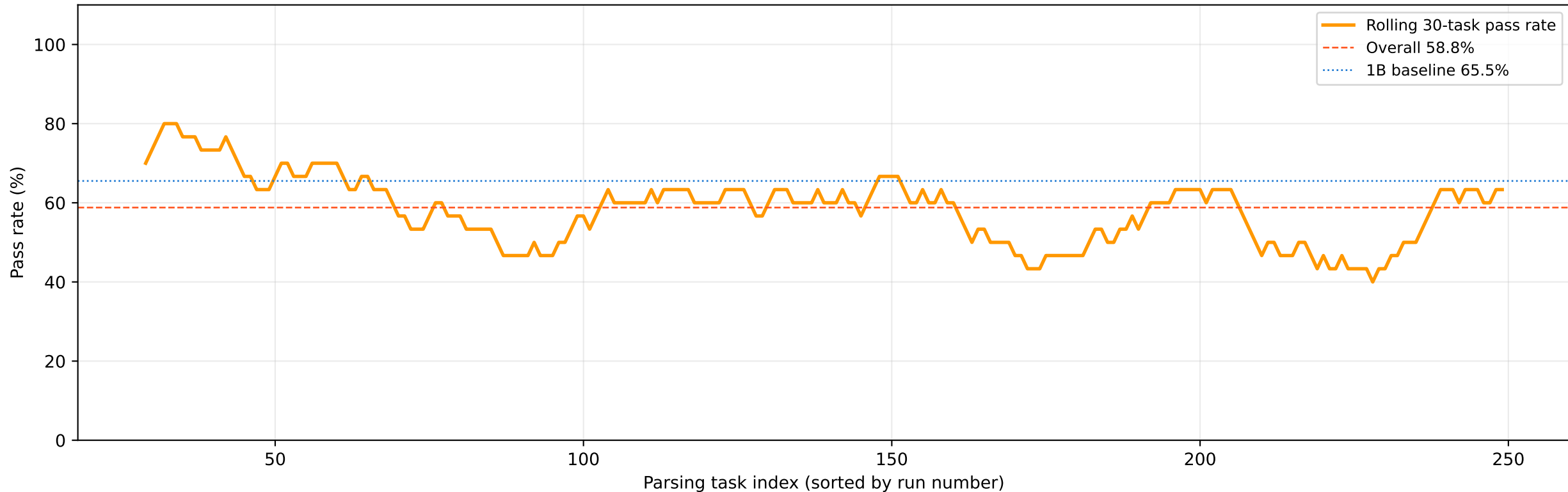
Parsing: Per-(Task-Type, Tribe) Pass Rate (1C)



Parsing: Tribe Routing Share Over Time



Parsing: Rolling Pass Rate Over Time (1C)



OPTION C — PARSING CASE STUDY

OVERVIEW

Phase 1C: 250 parsing tasks (58.8% pass)

Phase 1B: 58 parsing tasks (65.5% pass)

Regression: 6.7pp drop from 1B to 1C

PER-TASK-TYPE (Phase 1C)

roman_to_int : 0.0% pass top_exc=KeyError

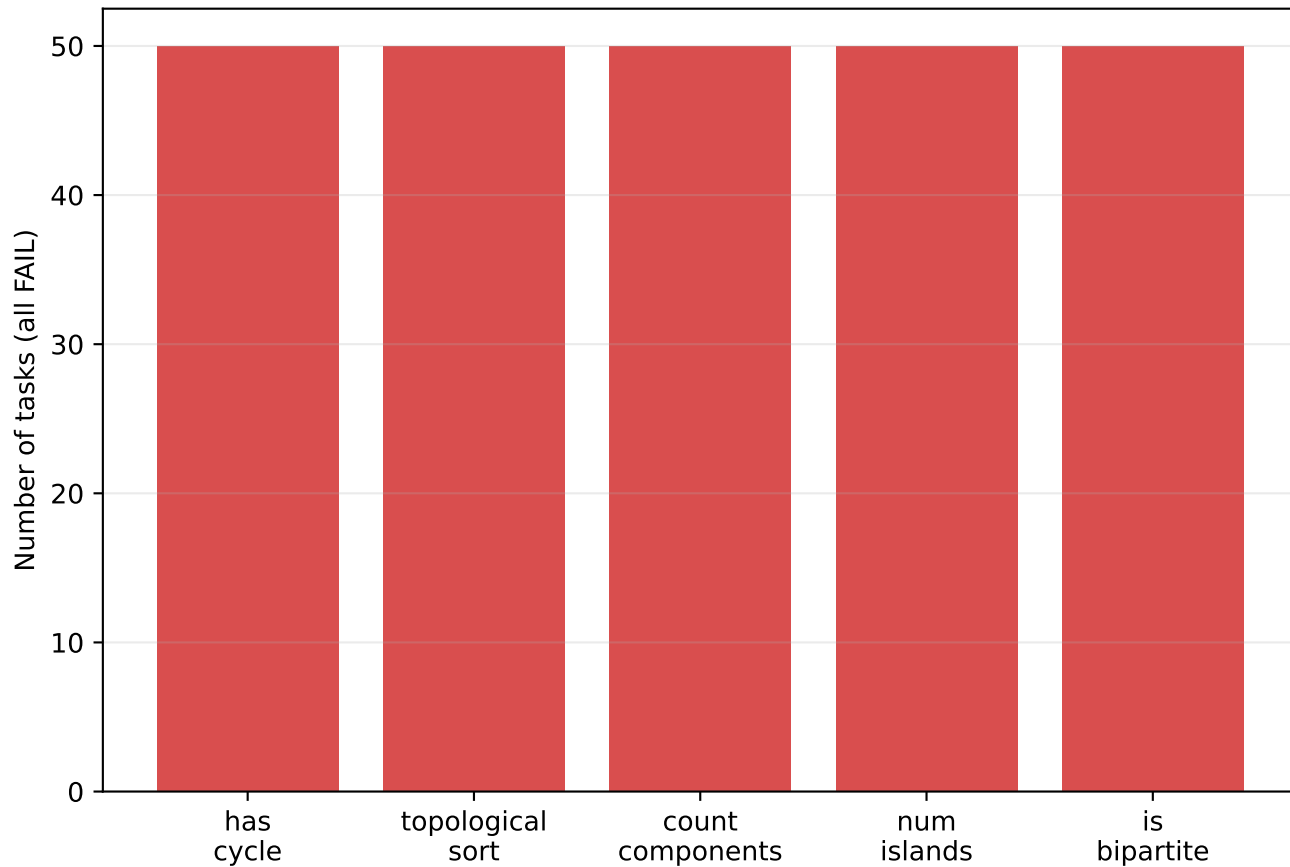
title_case : 100.0% pass top_exc=--

camel_to_snake : 94.0% pass top_exc=AssertionError

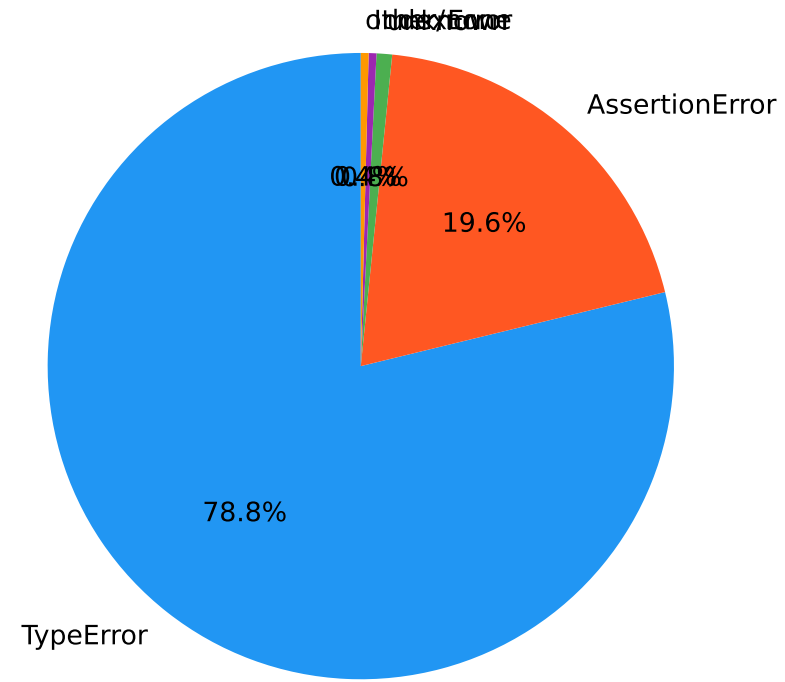
count_vowels : 100.0% pass top_exc=--

decode_run_length : 0.0% pass top_exc=ValueError

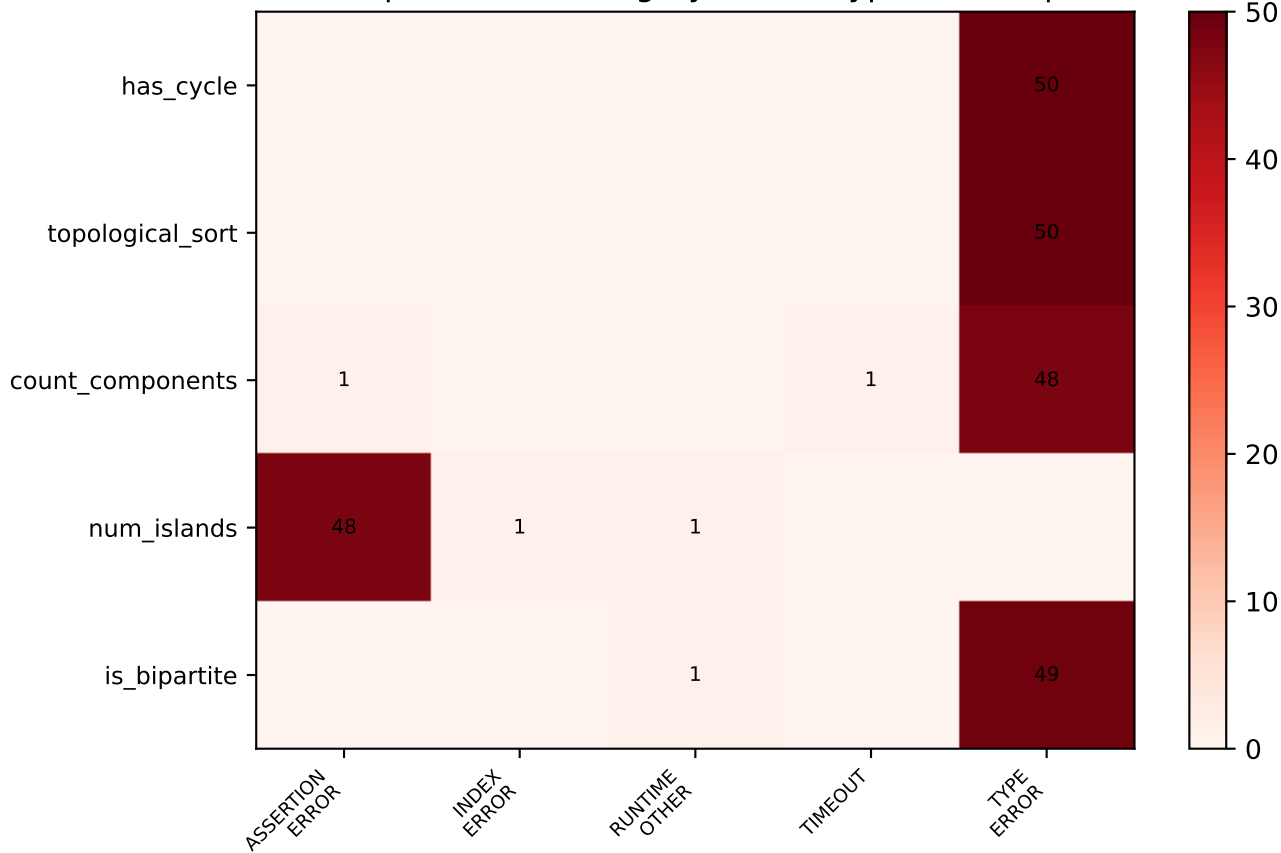
Graph: Per-Task-Type Distribution (1C, 0% pass)



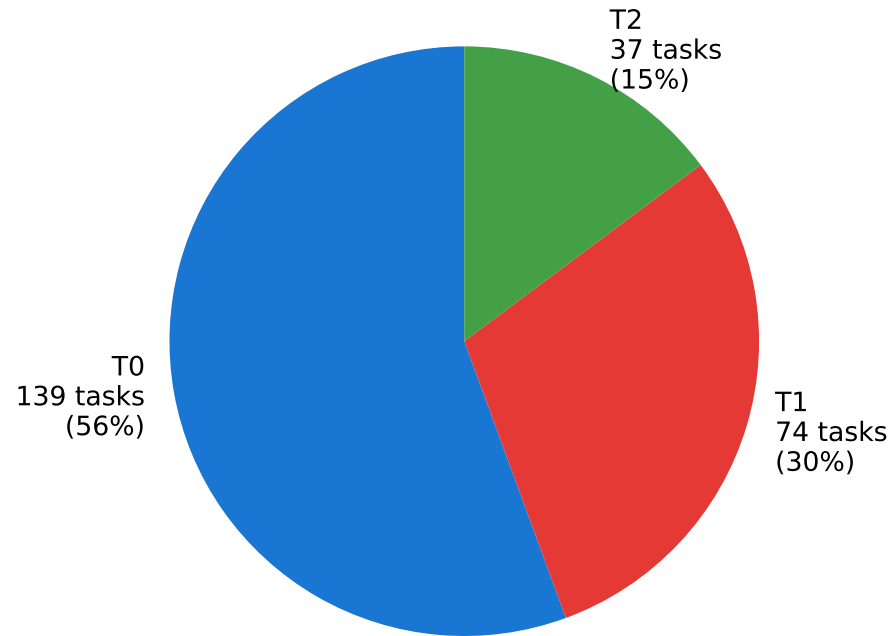
Graph Failure Modes (1C, 250 tasks, 0 passes)



Graph: Failure Category × Task Type Heatmap



Graph: Routing Distribution (all FAIL regardless)



OPTION D — GRAPH CEILING ANALYSIS

RESULT: 0% pass rate — confirmed across all 5 graph task types

SCOPE

5 graph task types: `has_cycle`, `topological_sort`, `count_components`,

`num_islands`, `is_bipartite`

Phase 1C: 250 tasks | Phase 1B: 50 tasks

0 passes across both phases — not a sampling artifact

FAILURE TAXONOMY

`TypeError` : 197 (78.8%)

`AssertionError` : 49 (19.6%)

`unknown` : 2 (0.8%)