Phase 1: Benchmark Setup + Baseline Results

Marko Jojic, Jaya Adithya Pavuluri, Han Nguyen, Mohit Jain, Chris Kurian

Setup Overview

Models & Configuration:

The τ-bench is a benchmark that evaluates tool-using conversational agents in realistic simulated user scenarios for *airline* and *retail* Domains. For Phase 1 of our assignment we tested the Qwen3 models 4B, 8B, 14B, 32B across the retail and airline domains. All experiments used Qwen3-32B as the fixed user agent.

- **Agent models:** Qwen3-4B/8B/14B/32B
- **User simulator:** Qwen3-32B
- **Infrastructure:** Sol cluster with 2× A100-80GB GPUs
- **Trials:** 5 per configuration needed for pass^k with k=1 to 5

We ran these experiments on the 2 domains:

- **Retail:** 115 tasks involving orders, refunds, product searches
- **Airline:** 50 tasks for flight booking and reservation management

Strategies Tested:

1. **ReAct**: Agent reasons explicitly before each action (Thought -> Action -> Observation loop)
2. **ACT**: Direct action selection without showing reasoning steps
3. **Function Calling (FC):** Structured JSON Function Calling with explicit schemas

The Work We did:

We were able to complete benchmarks for each combination of model, strategy, and dataset with 5 runs.

Infrastructure & Automation:

- Set up Sol cluster access and configured SLURM jobs with proper QoS settings and environments
- Designed multi-GPU allocation strategy for larger models

Experiment Execution:

- Ran complete evaluation suite: 4 model sizes × 3 strategies × 2 domains × 5 trials
- Built our own automation scripts to run experiments in a single interactive session(starting vLLM servers, running tau-bench for each strategy and cleanup between experiments)

Analysis:

- Wrote scripts to compute pass^k metrics from log files using the same formula referenced in the taubench paper

Contributions

Jay and Marko set up the environments as well as the scripts to run benchmarks. From here, we divided the runs across our group, which you can see in the names in the results table. We all contributed to the document, most of the work was in running the experiments.

Results

Below you can find the metrics for all Qwen3 models for the **Retail** and **Airline** domains.

Retail Results Table:

ACT						
Contribution	Agent Model	pass^1	pass^2	pass^3	pass^4	pass^5
Jay	Qwen 4B	0.1096	0.0478	0.0278	0.0191	0.0174
Chris	Qwen 8B	0.127	0.0583	0.0374	0.0261	0.0174
Marko	Qwen 14B	0.2017	0.087	0.0504	0.0348	0.0261
Marko	Qwen 32B	0.1583	0.0643	0.0322	0.0174	0.0087
REACT						
Contribution	Agent Model	pass^1	pass^2	pass^3	pass^4	pass^5
Jay	Qwen 4B	0.0922	0.0461	0.0304	0.0226	0.0174
Chirs	Qwen 8B	0.1556	0.0764	0.0389	0.0167	0
Marko	Qwen 14B	0.3096	0.1635	0.1052	0.073	0.0522
Marko	Qwen 32B	0.3374	0.2	0.1357	0.0991	0.0783
FC						
Contribution	Agent Model	pass^1	pass^2	pass^3	pass^4	pass^5
Jay	Qwen 4B	0.0974	0.0487	0.0313	0.0226	0.0174

Chris	Qwen 8B	0.1217	0.047	0.0235	0.0139	0.0087
Marko	Qwen 14B	0.2678	0.1339	0.0904	0.0713	0.0609
Marko	Qwen 32B	0.2817	0.1348	0.0696	0.0365	0.0174

Airline Results Table:

ACT						
Contribution	Agent Model	pass^1	pass^2	pass^3	pass^4	pass^5
Mohit	Qwen 4B	0.28	0.206	0.168	0.148	0.14
Han	Qwen 8B	0.204	0.128	0.09	0.068	0.06
Marko	Qwen 14B	0.244	0.156	0.114	0.092	0.08
Marko	Qwen 32B	0.308	0.186	0.12	0.08	0.06
REACT						
Contribution	Agent Model	pass^1	pass^2	pass^3	pass^4	pass^5
Mohit	Qwen 4B	0.296	0.228	0.19	0.164	0.14
Han	Qwen 8B	0.252	0.186	0.142	0.108	0.08
Marko	Qwen 14B	0.288	0.188	0.156	0.126	0.12
Marko	Qwen 32B	0.344	0.242	0.204	0.188	0.18
FC						
Contribution	Agent Model	pass^1	pass^2	pass^3	pass^4	pass^5
Mohit	Qwen 4B	0.248	0.136	0.074	0.032	0
Han	Qwen 8B	0.224	0.114	0.064	0.036	0.02
Marko	Qwen 14B	0.236	0.09	0.04	0.016	0
Marko	Qwen 32B	0.296	0.166	0.114	0.088	0.08

What we learned

Model scaling improves first-trial performance substantially* but only marginally improves reliability:

Bigger models do better on a single try: retail ReAct goes from 0.092 (4B) to 0.337 (32B), making it nearly 4 times better.

However, “better once” does not mean “reliably better" as you can see from the data in pass^5 for retail ReAct 4B and 32B. 4B reproduces its successes around 18.5%

(0.017/0.092) of the time, and 32B only around 23.2% (0.078/0.337). Shows that the models are barely more reliable despite being 8 times larger

Retail 8B ReAct (.156) scores higher than 4B (.092) on a single run but 8B hits pass^5 = 0.000. Meaning it never solves any task consistently. Again showing being bigger doesn't automatically make it more reliable

ReAct consistently stronger in larger scale models:

At 4B, ReAct and ACT are near-equivalent across both domains: retail (ACT: 0.110 vs ReAct: 0.092), airline (ACT: 0.280 vs ReAct: 0.296) It doesn't seem to matter which you pick at that small scale.

At 32B: ReAct *wins*, retail (0.337 vs 0.158), airline (0.344 vs 0.308) and even more significantly for pass^5

Practical takeaway: ReAct makes the model "think out loud" before acting. Small models don't have enough reasoning ability to benefit from this but large models do so the advantage only shows up at scale

- Use ACT for small/fast deployments
- use ReAct whenever you have a 14B+ model

Domain performance gap smaller with bigger models, but task-clustering divergence persists:

At 4B, airline is about 3 times easier than retail (0.296 vs 0.092 ReAct pass^1). By 32B they're almost tied (0.344 vs 0.337). This shows that bigger models close the *raw score gap* between the 2 domains.

But the *consistency gap* stays because airline 32B ReAct solves the same tasks correctly in all 5 runs 52% of the time (0.180/0.344) while retail only manages a 23% (0.078/0.337)

Practical takeaway: An airline agent fails more predictably, so this will be easier to identify and fix. However, A retail agent's failures are more unpredictable, it might handle a task fine today and fail on the same task tomorrow, which might be much harder to debug and fix.

Function Calling reliability collapse is domain and scale specific:

Function Calling (FC) looks reasonable on a single try as seen in airline 4B FC scores 0.248, airline 32B FC scores 0.296. FC here competes with with ReAct

However, when ran 5 times: airline 4B and 14B FC both hit 0.000, meaning not one task is ever solved reliably across all 5 runs

Retail FC stays non-zero at every model size, suggesting FC works fine for simpler, shorter tasks but breaks down when tasks require tracking many constraints across many steps (like airline bookings)

Note: As seen here sometimes it takes 5 runs to see the zeros for reliability. As mentioned before, this is exactly why single-trial evaluation is insufficient for real deployments.

Challenges We Hit

Initially we faced several challenges, but nothing that had us stuck for long. A bug in the latest ‘litellm’ version caused runs to terminate mid-experiment, this was only caught after inspecting incomplete output files and we fixed this by downgrading the package.

Upgrading vLLM to support Qwen3 models surfaced a numpy/PyTorch version conflict that needed manual resolution, and intermittent "connection closed" errors during long SLURM jobs were fixed by patching `tau_bench/envs/user.py` to reinitialize the HTTP client on failures.

GPU fairshare became a bottleneck later on in project, with some allocations arriving past 1 AM. We also switched from Qwen3-32B quantized to Qwen3-32B full as the user simulator after noticing hallucinated outputs were leading to agent failures and giving us false results, which needed 2 GPUs increasing the usage on our asu sol/ research computing accounts.

JSON Logs:
https://drive.google.com/drive/folders/19LPIwjcm35aoNj9RptA-xlX4tZjKKMLN?usp=sharing’

```
[mjojic@sol-login01:~/CSE598/tau-bench]$ ./run_analyze_all.sh
Pass^k analysis for ./results/act_retail_trials5_qwen_8b/act-Qwen3-8B-0.0_range_0--1_user-Qwen3-32B-llm_0130133152.json (k=5)
--------------------------------------------------
Total accuracy: 12.70% (73/575 task-runs)
pass^1: 0.1270 (over 115 tasks)
pass^2: 0.0583 (over 115 tasks)
pass^3: 0.0374 (over 115 tasks)
pass^4: 0.0261 (over 115 tasks)
pass^5: 0.0174 (over 115 tasks)
Pass^k analysis for ./results/act_airline_trials5_qwen_8b/act-Qwen3-8B-0.0_range_0--1_user-Qwen3-32B-llm_0130133550.json (k=5)
--------------------------------------------------
Total accuracy: 20.40% (51/250 task-runs)
pass^1: 0.2040 (over 50 tasks)
pass^2: 0.1280 (over 50 tasks)
pass^3: 0.0900 (over 50 tasks)
pass^4: 0.0680 (over 50 tasks)
pass^5: 0.0600 (over 50 tasks)
Pass^k analysis for ./results/tool-calling_airline_trials5_qwen_4b/tool-calling-Qwen3-4B-0.0_range_0--1_user-Qwen3-32B-llm_0210201313.json (k=5)
--------------------------------------------------
Total accuracy: 24.80% (62/250 task-runs)
pass^1: 0.2480 (over 50 tasks)
pass^2: 0.1360 (over 50 tasks)
pass^3: 0.0740 (over 50 tasks)
pass^4: 0.0320 (over 50 tasks)
pass^5: 0.0000 (over 50 tasks)
Pass^k analysis for ./results/tool-calling_retail_trials5_qwen_14b/tool-calling-Qwen3-14B-0.0_range_0--1_user-Qwen3-32B-llm_0201040648.json (k=5)
--------------------------------------------------
Total accuracy: 26.78% (154/575 task-runs)
pass^1: 0.2678 (over 115 tasks)
pass^2: 0.1339 (over 115 tasks)
pass^3: 0.0904 (over 115 tasks)
pass^4: 0.0713 (over 115 tasks)
pass^5: 0.0609 (over 115 tasks)
Pass^k analysis for ./results/react_airline_trials5_qwen_32b/react-Qwen3-32B-0.0_range_0--1_user-Qwen3-32B-llm_0201010843.json (k=5)
--------------------------------------------------
Total accuracy: 34.40% (86/250 task-runs)
pass^1: 0.3440 (over 50 tasks)
pass^2: 0.2420 (over 50 tasks)
pass^3: 0.2040 (over 50 tasks)
pass^4: 0.1880 (over 50 tasks)
pass^5: 0.1800 (over 50 tasks)
Pass^k analysis for ./results/react_retail_trials5_qwen_8b/react-Qwen3-8B-0.0_range_0--1_user-Qwen3-32B-llm_0131180350.json (k=5)
--------------------------------------------------
Total accuracy: 13.16% (70/532 task-runs)
pass^1: 0.1556 (over 72 tasks)
pass^2: 0.0764 (over 72 tasks)
pass^3: 0.0389 (over 72 tasks)
pass^4: 0.0167 (over 72 tasks)
pass^5: 0.0000 (over 72 tasks)
Pass^k analysis for ./results/act_airline_trials5_qwen_32b/act-Qwen3-32B-0.0_range_0--1_user-Qwen3-32B-llm_0201033030.json (k=5)
--------------------------------------------------
Total accuracy: 30.80% (77/250 task-runs)
pass^1: 0.3080 (over 50 tasks)
pass^2: 0.1860 (over 50 tasks)
pass^3: 0.1200 (over 50 tasks)
pass^4: 0.0800 (over 50 tasks)
pass^5: 0.0600 (over 50 tasks)
Pass^k analysis for ./results/react_retail_trials5_qwen_32b/react-Qwen3-32B-0.0_range_0--1_user-Qwen3-32B-llm_0201105215.json (k=5)
--------------------------------------------------
Total accuracy: 33.74% (194/575 task-runs)
pass^1: 0.3374 (over 115 tasks)
pass^2: 0.2000 (over 115 tasks)
pass^3: 0.1357 (over 115 tasks)
pass^4: 0.0991 (over 115 tasks)
pass^5: 0.0783 (over 115 tasks)
```

```
Pass^k analysis for ./results/act_retail_trials5_qwen_14b/act-Qwen3-14B-0.0_range_0--1_user-Qwen3-32B-llm_0201105240.json (k=5)
--------------------------------------------------
Total accuracy: 20.17% (116/575 task-runs)
pass^1: 0.2017 (over 115 tasks)
pass^2: 0.0870 (over 115 tasks)
pass^3: 0.0504 (over 115 tasks)
pass^4: 0.0348 (over 115 tasks)
pass^5: 0.0261 (over 115 tasks)
Pass^k analysis for ./results/tool-calling_retail_trials5_qwen_32b/tool-calling-Qwen3-32B-0.0_range_0--1_user-Qwen3-32B-llm_0201152610.json (k=5)
--------------------------------------------------
Total accuracy: 28.17% (162/575 task-runs)
pass^1: 0.2817 (over 115 tasks)
pass^2: 0.1348 (over 115 tasks)
pass^3: 0.0696 (over 115 tasks)
pass^4: 0.0365 (over 115 tasks)
pass^5: 0.0174 (over 115 tasks)
Pass^k analysis for ./results/react_airline_trials5_qwen_8b/react-Qwen3-8B-0.0_range_0--1_user-Qwen3-32B-llm_0131175900.json (k=5)
--------------------------------------------------
Total accuracy: 25.20% (63/250 task-runs)
pass^1: 0.2520 (over 50 tasks)
pass^2: 0.1860 (over 50 tasks)
pass^3: 0.1420 (over 50 tasks)
pass^4: 0.1080 (over 50 tasks)
pass^5: 0.0800 (over 50 tasks)
Pass^k analysis for ./results/tool-calling_airline_trials5_qwen_8b/tool-calling-Qwen3-8B-0.0_range_0--1_user-Qwen3-32B-llm_0130001030.json (k=5)
--------------------------------------------------
Total accuracy: 22.40% (56/250 task-runs)
pass^1: 0.2240 (over 50 tasks)
pass^2: 0.1140 (over 50 tasks)
pass^3: 0.0640 (over 50 tasks)
pass^4: 0.0360 (over 50 tasks)
pass^5: 0.0200 (over 50 tasks)
Pass^k analysis for ./results/tool-calling_retail_trials5_qwen_8b/tool-calling-Qwen3-8B-0.0_range_0--1_user-Qwen3-32B-llm_0130000731.json (k=5)
--------------------------------------------------
Total accuracy: 12.17% (70/575 task-runs)
pass^1: 0.1217 (over 115 tasks)
pass^2: 0.0470 (over 115 tasks)
pass^3: 0.0235 (over 115 tasks)
pass^4: 0.0139 (over 115 tasks)
pass^5: 0.0087 (over 115 tasks)
Pass^k analysis for ./results/act_retail_trials5_qwen_32b/act-Qwen3-32B-0.0_range_0--1_user-Qwen3-32B-llm_0201164647.json (k=5)
--------------------------------------------------
Total accuracy: 15.83% (91/575 task-runs)
pass^1: 0.1583 (over 115 tasks)
pass^2: 0.0643 (over 115 tasks)
pass^3: 0.0322 (over 115 tasks)
pass^4: 0.0174 (over 115 tasks)
pass^5: 0.0087 (over 115 tasks)
Pass^k analysis for ./results/tool-calling_airline_trials5_qwen_32b/tool-calling-Qwen3-32B-0.0_range_0--1_user-Qwen3-32B-llm_0201064103.json (k=5)
--------------------------------------------------
Total accuracy: 29.60% (74/250 task-runs)
pass^1: 0.2960 (over 50 tasks)
pass^2: 0.1660 (over 50 tasks)
pass^3: 0.1140 (over 50 tasks)
pass^4: 0.0880 (over 50 tasks)
pass^5: 0.0800 (over 50 tasks)
Pass^k analysis for ./results/react_retail_trials5_qwen_14b/react-Qwen3-14B-0.0_range_0--1_user-Qwen3-32B-llm_0131203002.json (k=5)
--------------------------------------------------
Total accuracy: 30.96% (178/575 task-runs)
pass^1: 0.3096 (over 115 tasks)
pass^2: 0.1635 (over 115 tasks)
pass^3: 0.1052 (over 115 tasks)
pass^4: 0.0730 (over 115 tasks)
pass^5: 0.0522 (over 115 tasks)
```

```
Pass^k analysis for ./results/react_airline_trials5_qwen_14b/react-Qwen3-14B-0.0_range_0--1_user-Qwen3-32B-llm_0131211644.json (k=5)
--------------------------------------------------
Total accuracy: 28.80% (72/250 task-runs)
pass^1: 0.2880 (over 50 tasks)
pass^2: 0.1880 (over 50 tasks)
pass^3: 0.1560 (over 50 tasks)
pass^4: 0.1360 (over 50 tasks)
pass^5: 0.1200 (over 50 tasks)
Pass^k analysis for ./results/tool-calling_airline_trials5_qwen_14b/tool-calling-Qwen3-14B-0.0_range_0--1_user-Qwen3-32B-llm_0201005750.json (k=5)
--------------------------------------------------
Total accuracy: 23.60% (59/250 task-runs)
pass^1: 0.2360 (over 50 tasks)
pass^2: 0.0900 (over 50 tasks)
pass^3: 0.0400 (over 50 tasks)
pass^4: 0.0160 (over 50 tasks)
pass^5: 0.0000 (over 50 tasks)
Pass^k analysis for ./results/act_airline_trials5_qwen_14b/act-Qwen3-14B-0.0_range_0--1_user-Qwen3-32B-llm_0131233237.json (k=5)
--------------------------------------------------
Total accuracy: 24.40% (61/250 task-runs)
pass^1: 0.2440 (over 50 tasks)
pass^2: 0.1560 (over 50 tasks)
pass^3: 0.1140 (over 50 tasks)
pass^4: 0.0920 (over 50 tasks)
pass^5: 0.0800 (over 50 tasks)
```

```
[jpavulu2@sol-login01:~/agentic_ai/temp]$ python3 analyze_passk.py
Pass^k analysis for retail_act-Qwen3-4B-0.0_range_0--1_user-Qwen3-32B-llm_0201221019.json (k=5)
Total accuracy: 10.96% (63/575 task-runs)
pass^1: 0.0696 (over 115 tasks)
pass^2: 0.1043 (over 115 tasks)
pass^3: 0.1130 (over 115 tasks)
pass^4: 0.1391 (over 115 tasks)
pass^5: 0.1217 (over 115 tasks)

Pass^k analysis for retail_react-Qwen3-4B-0.0_range_0--1_user-Qwen3-32B-llm_0202040414.json (k=5)
Total accuracy: 9.22% (53/575 task-runs)
pass^1: 0.1043 (over 115 tasks)
pass^2: 0.0870 (over 115 tasks)
pass^3: 0.0783 (over 115 tasks)
pass^4: 0.0957 (over 115 tasks)
pass^5: 0.0957 (over 115 tasks)

Pass^k analysis for tool-calling-Qwen3-4B-0.0_range_0--1_user-Qwen3-32B-llm_0201111240 (1).json (k=5)
Total accuracy: 9.74% (56/575 task-runs)
pass^1: 0.0870 (over 115 tasks)
pass^2: 0.1130 (over 115 tasks)
pass^3: 0.0957 (over 115 tasks)
pass^4: 0.1043 (over 115 tasks)
pass^5: 0.0870 (over 115 tasks)

[jpavulu2@sol-login01:~/agentic_ai/temp]$
```

```
[mjojic@sol-login01:~/CSE598/tau-bench]$ python analyze_pass_k.py
Pass^k analysis for /home/mjojic/CSE598/tau-bench/results/airline-Act-Qwen3-4B-llm_0202210600.json (k=5)
--------------------------------------------------
Total accuracy: 28.00% (70/250 task-runs)
pass^1: 0.2800 (over 50 tasks)
pass^2: 0.2060 (over 50 tasks)
pass^3: 0.1680 (over 50 tasks)
pass^4: 0.1480 (over 50 tasks)
pass^5: 0.1400 (over 50 tasks)
[mjojic@sol-login01:~/CSE598/tau-bench]$ python analyze_pass_k.py
Pass^k analysis for /home/mjojic/CSE598/tau-bench/results/airline-React-Qwen3-4B-llm_0203154512.json (k=5)
--------------------------------------------------
Total accuracy: 29.60% (74/250 task-runs)
pass^1: 0.2960 (over 50 tasks)
pass^2: 0.2280 (over 50 tasks)
pass^3: 0.1900 (over 50 tasks)
pass^4: 0.1640 (over 50 tasks)
pass^5: 0.1400 (over 50 tasks)
```