Construct Validity Failures in Agentic AI Benchmarks

An Empirical Audit · **Himaneesh Sompalle** · Stonehill International School, Bengaluru, India

THE CORE PROBLEM

One score cannot summarize an agent.

Agent leaderboards rank models on a single number. But five benchmarks — **τ-Retail, τ-Airline, SWE-bench Verified, GPQA Diamond, MMLU-Pro** — across **15 frontier & mid-tier models** do not induce a stable ranking.

0.67
Mean cross-benchmark ρ · range 0.10–0.92

0.10
τ-Airline × MMLU-Pro — effectively zero

22%
of model pairs invert rank — up to 40%

0
rank-spread — Claude Sonnet 4.5 (1st) & Opus 4 (3rd)

FIGURE 1

Cross-benchmark rank correlations

Spearman ρ across 15 models. Two latent clusters emerge, with τ-Airline carrying the discriminant signal.

	τ-Ret	τ-Air	SWE	GPQA	MMLU
τ-Retail	1.00	.75	.81	.76	.80
τ-Airline	.75	1.00	.66	.49	.10
SWE-bench	.81	.66	1.00	.72	.69
GPQA	.76	.49	.72	1.00	.92
MMLU-Pro	.80	.10	.69	.92	1.00
	TOOL-USE · mean ρ 0.74			REASONING · ρ 0.92	

- **τ-Airline × MMLU-Pro = 0.10** — two "agent" benchmarks, essentially uncorrelated.
- Even **τ-Retail & τ-Airline** — same framework — correlate at only **0.75**.

Reasoning specialists **o1, o3-mini, o4-mini** place top-5 on GPQA / MMLU-Pro but fall to **6th–12th** on tool-use — rank spreads of 7–9. Yet the constructs are not anti-correlated: **Claude Sonnet 4.5** ranks 1st across the board.

FIGURE 3 · RANK PROFILES

No stable ranking exists

Each line traces one model's rank (1 = best). Reasoning specialists dive on tool-use; Sonnet 4.5 holds rank 1.

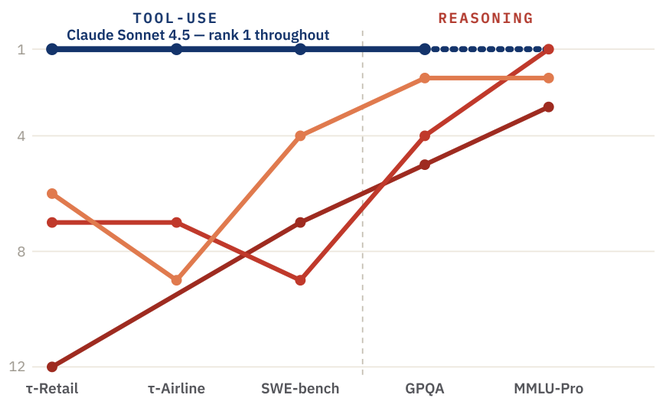

TABLE 6

Desiderata scorecard

✓ satisfied · ~ partial · ✗ not satisfied. None meet all five.

	D1	D2	D3	D4	D5
τ-Retail	✓	✗	~	✓	✗
τ-Airline	✓	✗	~	✓	✗
SWE-bench	✓	✗	✗	~	✗
GPQA	✓	✗	✗	✗	✗
MMLU-Pro	~	✗	✗	~	✗

THE PRESCRIPTION

Five desiderata for construct-valid evaluation

D1 — Explicit construct definition
State exactly what the benchmark measures — and what it does not.

D2 — Convergent validity evidence
Show ρ > 0.80 with other benchmarks of the same construct.

D3 — Discriminant validity evidence
Show low correlation with benchmarks of a different construct.

D4 — Per-construct score reporting
Publish capability profiles, not a single composite number.

D5 — Ranking stability disclosure
Report the expected inversion rate against related benchmarks.

What practitioners should do

✓ Match the benchmark to your deployment context — a support agent? Weight τ-bench. A coding assistant? Weight SWE-bench.

✗ Never select models from composite leaderboards that average across unrelated constructs.

Himaneesh Sompalle · Stonehill International School, Bengaluru
himaneeshsompalle7@gmail.com

KDD '26 Workshop on Evaluation & Trustworthiness of Agentic AI · Jeju, Korea · August 2026

PAPER + CODE