Selecting LLM Judges for Agent Evaluation Pipelines:

Accuracy Masking and Review-Burden Tradeoffs

Punitha Ponnuraj
Independent Researcher · Plano, Texas, USA
punitha.p.raj@gmail.com

★ **Task-success accuracy stays tightly clustered across LLM judges — while side-effect accuracy diverges by 45–50 percentage points on the very same judges.**

1 Why This Matters

- LLM judges decide which agent trajectories get reviewed, filtered, or passed downstream.
- Judge selection typically relies on one number: headline task-success accuracy.
- A judge that looks reliable on task success can still fail at safety monitoring or review triage.

2 Accuracy Masking

Same 5 judges, same trajectories — very different spreads.

5 judge conditions

Task-success accuracy	Side-effect accuracy
1.6–4.4pp	**45.0–50.0pp**

Same headline accuracy, different underlying reliability.

3 Experimental Setup

Fixed	Expert majority-vote labels; Fisher's exact test for error-rate comparisons.
Varied	3 evaluated agents × 5 judge conditions × 3 benchmarks.
Data	WebArena, WorkArena, VisualWebArena — via AgentRewardBench (Lù et al., 2025).

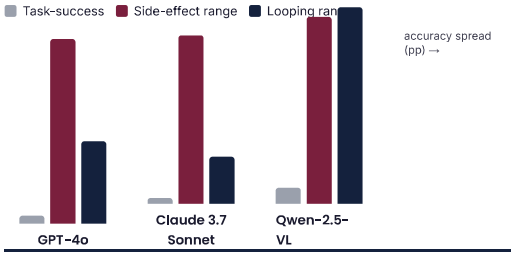
4 Pipeline Decision Framework

Pipeline decision	Relevant property
Benchmark scoring	Task-success accuracy
Side-effect monitoring	Recall / FNR
Human-review triage	Recall vs. review burden
Target-agent rollout	Profile stability

Task-success accuracy is directly relevant to only 1 of 6 common pipeline decisions.

5 Main Result: Accuracy Masking Persists

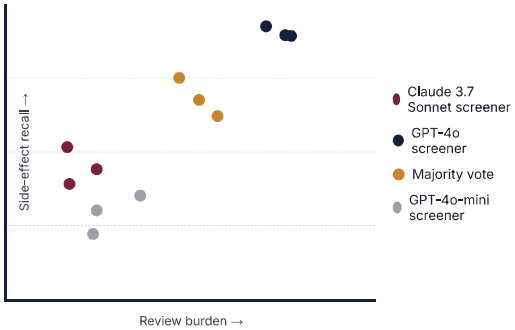
Task-success accuracy stays flat across agents; side-effect and looping accuracy do not.



Task-success spread	Side-effect spread	Looping spread
1.6–4.4pp	**45.0–50.0pp**	**12.6–52.5pp**

6 Review-Burden Tradeoffs

Each point is one escalation strategy across the 3 evaluated agents.



Claude 3.7 Sonnet gives the best recall-per-burden ratio among low-burden strategies on every agent; GPT-4o screening maximizes recall (90–94%) at 71–78% review burden.

7 Side-Effect Failure Profiles

Judge	Profile	FNR	FPR	Stable?
GPT-4o-mini	Silent misser	.64–.77	.24–.37	Yes
GPT-4o	High-recall over-flagger	.06–.09	.70–.77	Yes
Claude 3.7 Sonnet	Agent-dependent	.47–.59	.15–.24	No
Llama-3.3-70B	High-recall over-flagger	.23–.28	.65–.74	Yes
Qwen-2.5-VL	Inconclusive	.41–.50	.34–.49	Yes

FNR: false-negative rate on side-effect-positive trajectories. FPR: false-positive rate on side-effect-negative trajectories. Ranges span the 3 evaluated agents.

8 Behavior Depends on the Agent

- On GPT-4o agent trajectories, Claude 3.7 Sonnet is relatively balanced: FNR 0.47, FPR 0.15.
- On Claude and Qwen agent trajectories, it shifts miss-prone: FNR rises to 0.53–0.59.

A profile observed on one agent's trajectories may not transfer when the evaluated agent changes.

9 Looping: A Larger Positive Slice

115–205 loop-positive cases per agent vs. 17–32 for side effects — more statistical power.

GPT-4o-mini, loop-positive vs. overall accuracy
0.248 vs. 0.648
Δ = −40.0pp, $p < 0.001$

GPT-4o, Claude 3.7 Sonnet, and Llama-3.3-70B show the opposite pattern: better on loop-positive, worse on loop-negative trajectories.

10 Predictors of Judge Error

- **6.02×** odds of side-effect error for long trajectories (≥30 steps)
- **9.40×** odds of looping error for long trajectories
- **3.53×** GPT-4o-mini's odds of looping error — largest among judges

Trajectory length is an independent, usable escalation signal.

11 Practical Guidance

- ☑ Benchmark scoring → task-success + cost/latency
- ☑ High-recall SE monitoring → GPT-4o or Llama-3.3-70B
- ☑ Low-FP monitoring → Claude 3.7 Sonnet, audited
- ☑ Limited-budget triage → Claude 3.7 Sonnet screener
- ☑ Higher-budget triage → majority vote / GPT-4o
- ☑ Reliability monitoring → avoid GPT-4o-mini alone
- ☑ Target-agent rollout → re-audit on new trajectories

12 Takeaway

★ **Task-success accuracy is an incomplete criterion for judge selection. Combine it with dimension-level error profiles, target-agent behavior, and review budget.**

Code: github.com/ppon1086/accuracy-masking-agent-evaluation
Data: AgentRewardBench (Lù et al., 2025) via Hugging Face