

PayerBench

Claims reasoning, evaluation and post-training

A benchmark and tool-based RL environment built from de-identified outpatient claims, with evidence tools and actual insurer remittances as grading targets.

Yash Joshi & Roshan Patel / Arrow / October 22, 2026

24,302

claims in the source corpus

486 payers / five specialties

18,445

distinct training task IDs

12,618 observed + 5,827 constructed

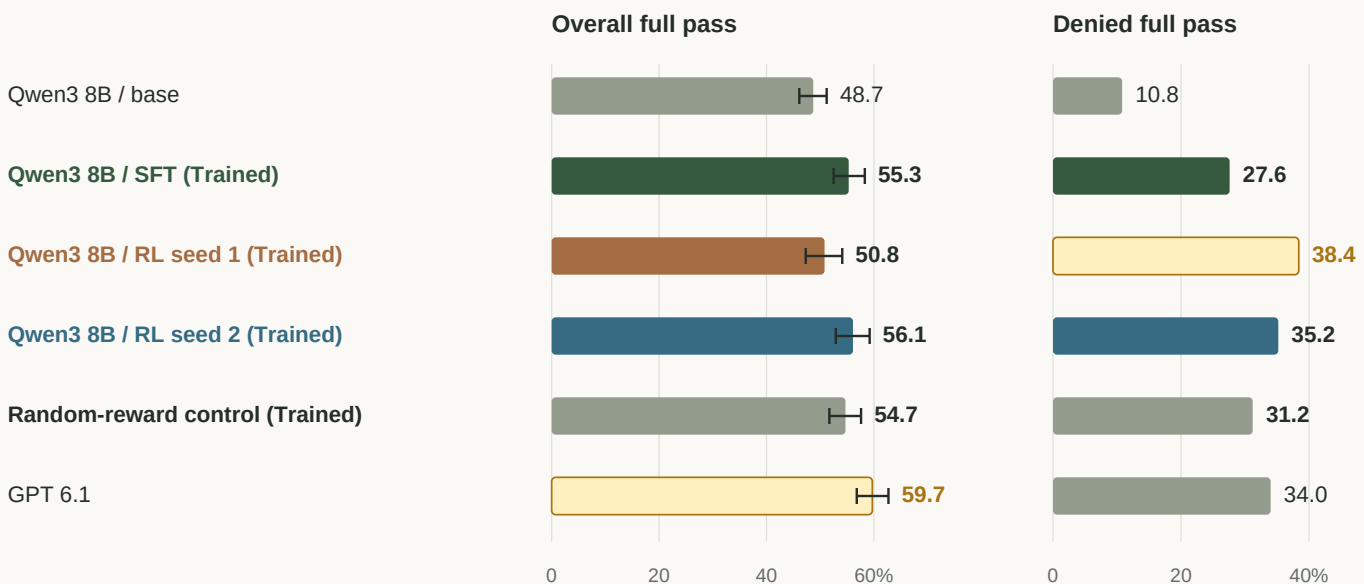
750

held-out claims evaluated

95 payers / three specialties

What post-training changed

Same 750 claims. Overall whiskers: 95% intervals. Denied bars: point estimates.



Five 8B policies and the highest-scoring reference are shown above. All 16 completed configurations appear on page 3; the complete evaluation scorecard is on page 2. Gold marks the highest observed value in each metric; it does not establish statistical superiority.

SFT: +6.7 pp overall

95% CI [4.4, 9.1] versus base. Denied full pass rises from 10.8% to 27.6%; paid full pass stays at 94.8%.

RL: +7.6 pp on denied claims

Seed 2 versus SFT. Overall gain: +0.8 pp, CI [-1.2, 2.8]. Paid full pass falls to 86.0%. Both RL seeds start independently from the same SFT checkpoint.

Evaluation scorecard

Production-derived claims, nine evidence tools and a deterministic remittance-based verifier. Primary comparison: RL seed 2 versus the same SFT checkpoint on 750 held-out claims.

59.7%

Frontier baseline / GPT 6.1

95% CI [56.8, 62.7]

+0.8 pp

RL versus SFT

95% CI [-1.2, 2.8]

+7.5 pp

SFT + RL versus base

95% CI [4.7, 10.1]

Metric / evidence status	Result and measurement scope
Frontier model baseline MEASURED	59.7% pass@1 GPT 6.1; 95% CI [56.8, 62.7]. 250 paid, 250 partial, 250 denied claims; 95 payers.
Held-out capability gain after RL MEASURED	+0.8 pp overall; +7.6 pp on denied claims Seed 2 vs SFT. Overall CI [-1.2, 2.8]; denied CI [4.4, 11.2]. Seed 1 overall: -4.5 pp.
Unseen task-family gain UNMEASURED	Not measured All six controlled families appear in training; no held-out-family experiment.
Production-derived holdout gain MEASURED	+0.8 pp incremental RL; +7.5 pp for SFT + RL Same holdout, not an independent replication. Pipeline CI [4.7, 10.1]; incremental RL CI includes zero.
Sim2Real transfer efficiency NOT APPLICABLE	Not applicable Production-derived training and evaluation; no separate simulator-to-production stage.
Human / reward agreement UNMEASURED	Not measured No independent expert agreement study. Labels are observed payer responses.
Reward false-positive rate UNMEASURED	Not established No independently adjudicated reward-correctness set; fixed attacks do not estimate this rate.
Reward exploit rate after red-teaming SCOPED	0 / 2,247 fixed attacks rewarded (0%) Three fixed mutation classes x 749 accepted references; no open-ended adversarial search.
Effective environment scale SCOPED	18,445 training task IDs / 6 controlled families 12,618 observed + 5,827 constructed tasks. Distinct IDs verified; semantic redundancy unaudited.
Saturation point SCOPED	Not established; 7,752 rollouts per RL seed observed Ten updates per seed; seed 2 dev peaks at step 15, then falls at 20. No saturation study.
Throughput SCOPED	2.83 rollout attempts/sec 8B / H100: 7,752 attempts / 2,738 rollout seconds; 0.42/sec with updates. Excludes validation/setup.
Environment reliability SCOPED	99.6% valid evaluated completions (1,550 / 1,556) Seed 2 after retries: prediction, process and integrity pass. Includes model behavior; not uptime.
Cost per 1M valid episodes SCOPED	Not measured; approximately \$93 reported experiment spend 29 GPU-hours for this version's SFT, RL, control and evaluation. No validated per-million unit cost.
Learning efficiency per \$10K training SCOPED	Not established; 3,876 rollouts to the selected RL checkpoint Five updates; 2.72 GPU-hours to selected seed 2 checkpoint, excluding SFT/validation/setup. No \$10K extrapolation.
Synthetic substitution gap UNMEASURED	Not measured No matched training comparison against a strongest cheap synthetic replica.

Measured: direct evaluation. **Scoped:** bounded audit, proxy or cost context. **Unmeasured:** evidence not yet established. **Not applicable:** outside this evaluation's design.

16 runs. One shared test.

Class-balanced full-prediction pass@1 on 750 claims: 250 paid, 250 partial and 250 denied. A pass requires the claim class, graded service-line outcomes and reason-code numbers, plus process and prediction-integrity checks.

Percentages throughout. Overall scores include 95% claim-bootstrap intervals.

Model / checkpoint	Overall / CI	Paid	Partial	Denied	Paid false alarms
GPT 6.1 Reference model	59.7% [56.8, 62.7]	93.6%	51.6%	34.0%	6.4%
GPT 6.1 / checklist prompt Prompting variant - no weight updates	59.3% [56.4, 62.4]	92.0%	51.2%	34.8%	8.0%
Claude Opus 5.5 Reference model	57.2% [54.3, 60.0]	94.8%	49.6%	27.2%	5.2%
Kimi K3 Reference model	56.3% [53.3, 59.1]	93.2%	47.6%	28.0%	6.0%
Qwen3 8B / RL seed 2 (Trained) Trained checkpoint	56.1% [52.9, 59.2]	86.0%	47.2%	35.2%	14.0%
GLM 5.3 Reference model	55.9% [52.9, 58.8]	93.2%	48.8%	25.6%	6.0%
Qwen3 8B / SFT (Trained) Trained checkpoint	55.3% [52.5, 58.3]	94.8%	43.6%	27.6%	5.2%
Random-reward control (Trained) Trained control / random reward	54.7% [51.7, 57.6]	92.0%	40.8%	31.2%	8.0%
Gemini 3.1 Pro Reference model	53.6% [50.7, 56.5]	88.0%	48.4%	24.4%	6.8%
DeepSeek V3.2 Reference model	53.1% [50.0, 56.0]	84.8%	48.0%	26.4%	14.8%
Qwen3 8B / RL seed 1 (Trained) Trained checkpoint	50.8% [47.3, 54.1]	66.0%	48.0%	38.4%	34.0%
Qwen3 235B A22B Reference model	50.0% [46.8, 53.2]	77.2%	46.4%	26.4%	14.8%
Gradient-boosted tree Statistical baseline / class-balanced tuning	49.5% [46.9, 51.9]	97.6%	39.2%	11.6%	2.0%
Qwen3 8B / base Base checkpoint	48.7% [46.1, 51.2]	94.8%	40.4%	10.8%	3.2%
gpt-oss-120b Reference model	29.6% [26.8, 32.4]	57.2%	24.4%	7.2%	30.4%
Qwen3 32B Reference model	29.1% [26.3, 31.9]	58.0%	21.6%	7.6%	41.2%

Gold: best observed value in each column, including ties. Lower paid false alarms is better. False alarms count actually paid claims classified as partial or denied.

Trained: 8B checkpoints post-trained in these experiments. GPT 6.1 with a checklist is a prompting variant. All 8B policies use greedy decoding in non-thinking mode; reference models use their evaluated provider settings. Intervals reflect sampled claims, not run-to-run variation.

Data and environment

Each episode asks an agent to inspect evidence and predict an insurer's response. The response is withheld from evidence tools and used as the grading target.

SOURCE CORPUS

24,302 claims / 47,613 lines

486 payers / five outpatient specialties

Names and direct identifiers are removed or replaced; dates and demographic details are transformed while preserving task relationships.

EVALUATED POPULATION

750 claims / 95 payers

558 allergy / immunology
165 dermatology
27 orthopedics

Training: 12,618 observed claims + 5,827 constructed tasks across six controlled rule families (18,445 distinct task IDs; semantic redundancy is unaudited). **Development:** 2,139 observed claims. **Later test period:** 9,545 claims, of which 750 are evaluated here. The later period starts June 1, 2026; train and dev are separated by patient, while test patients may have earlier training claims.

Inspect evidence. Commit one prediction. Receive a verdict.



Nine evidence tools: claim form, eligibility, authorizations, provider registry, code definitions, CMS NCCI edits, patient history, payer rules and similar training claims. The 12-call budget includes the terminal prediction. Payer rules and similar claims use the training split and exclude the current patient.

Measured operating context

Rollout throughput	2.83 attempts/sec 8B / H100; 0.42/sec including updates. Excludes validation and setup.
Valid evaluated completion	99.6% / 1,550 of 1,556 Seed 2 after retries; prediction, process and integrity pass. Not uptime.
Targeted verifier audit	0 of 2,247 attacks rewarded Three fixed mutation classes on 749 accepted references; not open-ended red-teaming.
Reported compute context	29 GPU-hours / about \$93 This version's SFT, RL, control and evaluation; not a per-episode unit cost.
Sim2Real transfer efficiency	Not applicable No separate simulator-to-production transfer stage in this setup.

Measurement limits: human/reward agreement, independent reward false positives, unseen-family transfer, synthetic substitution, saturation and normalized training cost are not established. The trained policies predate the integrity rescore; saved predictions were regraded without new inference. Observed payer responses are targets, not expert judgments of clinical or contractual correctness.