openauditor

Training and evaluation of a Qwen3-0.6B decision model with internal graph message passing.

Give openauditor a state and typed questions. It returns candidate probabilities, ordinal scores and binary probabilities instead of generated text. This report records its training, task accuracy, calibration, graph ablations, latency and closed-loop demonstrations.

Evidence	What was measured
Our decision tasks	23 task families; 6,217 test/heldout decisions, including reused regression cases
Our RL environments	768 episode evaluations across route, workflow and control; common seeds reused across interventions
Real demos	Structured-state ViZDoom and bounded live Wikipedia link navigation
Published Jev examples	408 typed questions on 20 selected cases; no paid API calls
Mechanism	Full graph, no graph, no messages, reversed edges on the same weights

The trained checkpoint

Selected stage: **environment_rl**. Native RL ran 90 updates; 18 had nonzero rollout-return variation. Exported context guard: 16,384 tokens; training guard: 4,096. An expanded input guard does not prove long-context reliability.

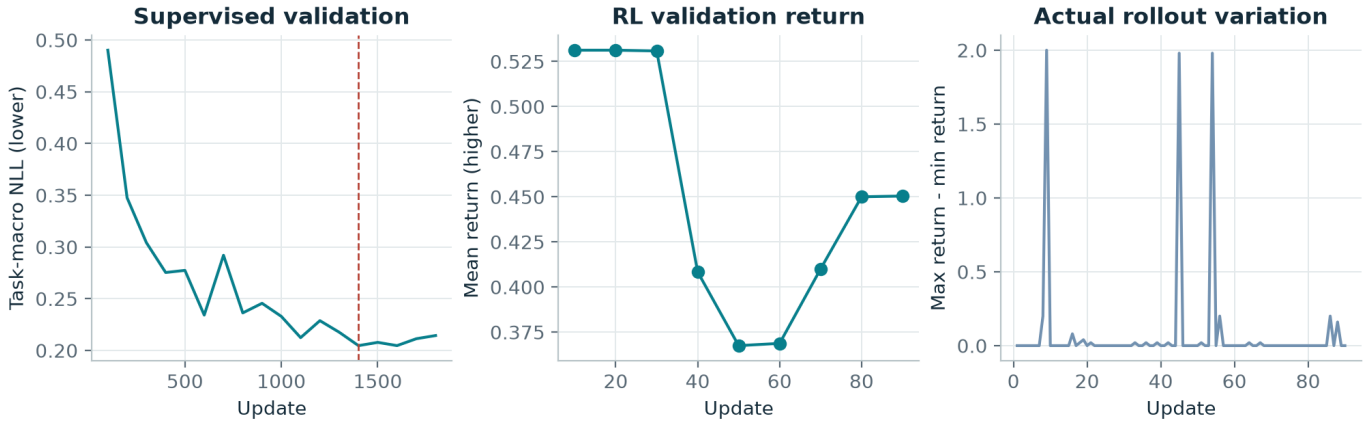
Results in brief

Fresh heldout evidence checks: 100.0%. Conditional-rule paraphrases: 78.3%. Reused synthetic security test: 44.0%. These task distributions measure different capabilities. The published-example comparison is limited to selected cases; it does not establish general Jev parity or hosted-service speed.

Repository and weights remain private. Converted session data is not redistributed. The model download is one merged checkpoint; optimizer states are separate training artifacts.

How openauditor was trained

Supervised contrast learning, then native policy gradients and validation-only calibration.



Prepared training decisions: **28,234**. Completed supervised updates: 1,800; best validation step: 1,400 (22,400 draws). Trainable parameters: 11,309,824. The run continued the initialized training adapter. Only LoRA and internal graph parameters are optimized; the red dashed line marks the selected supervised step.

State + typed question

Qwen layers 1-9

Internal graph bridge

Qwen layers 10-18

Internal graph bridge

Qwen layers 19-28 + logits

Inside each bridge

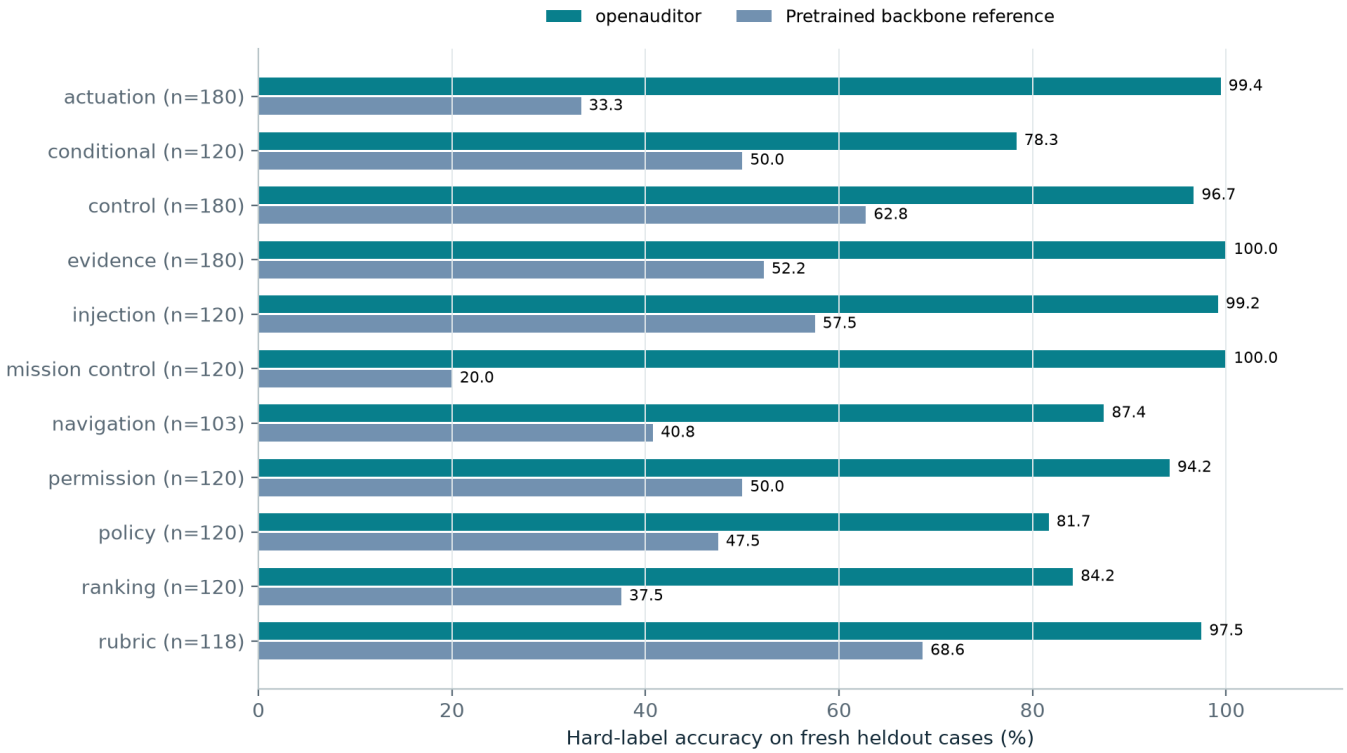
16 learned latent nodes, width 128.
Read prompt representations.
Learn top-4 incoming neighbors.
Run 2 message-passing rounds.
Write into the final decision token.

Later Qwen layers consume the graph-modified hidden state. These slots are not verified knowledge-graph entities.

Original verifier labels cover permission/negation, conditional rules, ranking, evidence, rubrics, count-based probabilities, control, actuation, navigation, policy and untrusted notes. Control-shaped training is new in this attempt; the Doom test is not a zero-exposure claim. No Jev examples or outputs were used for training or calibration.

Task accuracy

New cases and instruction paraphrases; not a claim that every family is unseen.



The chart uses only hard-label decisions in the fresh heldout contrast families. Ambiguous multi-action targets and probability-count questions are excluded from hard-label accuracy. Split grouping keeps paired instructions and identical states together. Larger graphs and selected paraphrases appear in heldout; some other families are IID cases.

The reference uses the pinned pretrained Qwen with zero LoRA/graph writes and the same validation-only calibration partition. It is an evaluation reference, not another product checkpoint or an equal-budget graph-free training experiment.

Probability-count task: mean absolute probability error 0.1482 over 120 heldout questions. These targets describe uniform sampling from known counts, not confidence in arbitrary real-world judgments.

All fresh and regression heldout data combined: 2,351 decisions. Per-task NLL, Brier diagnostics and group-level consistency are in the accompanying JSON. A soft-target accuracy field measures target mass, so it is not always ordinary accuracy.

Regression results

These old pilot test cases were already inspected; call them regression diagnostics, not fresh evidence.

Task	Cases	openauditor
CLINC routing	150	97.3%
Emotion	150	89.3%
CommonsenseQA	150	54.7%
Threshold rule	150	100.0%
Ordinal rubric	150	100.0%
WinoGrande	150	58.0%
Synthetic agent trace	150	100.0%
Synthetic security policy	150	44.0%

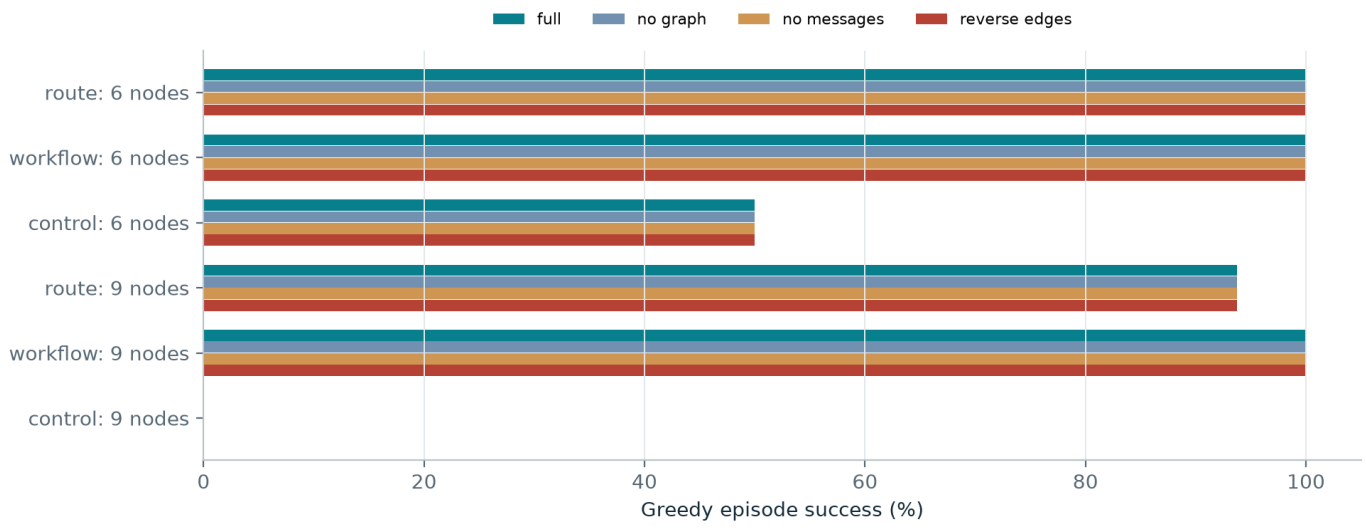
Local-session task: reported execution outcome

Negative-class (failure) recall: **4.2%**. Always-success accuracy: 89.0%. Failure AUROC: 0.764. A high headline accuracy can hide a model that never flags failures; use these diagnostics and the full confusion matrix.

This target is reported tool execution success, not authorization, correctness of an entire coding task, or successful completion of a user's goal. Session redaction does not prove de-identification, and private data prevents exact public reproduction of this mixed pilot.

Closed-loop task success

Fresh common seeds across interventions; the same checkpoint and action policy throughout.



Each bar uses 16 episodes. Training used 6-node difficulty; evaluation also uses 9-node difficulty. Seed offset: 30000000. Greedy results match API argmax; stochastic results and forecasts are reported separately in JSON.

Control transfer failure: greedy success is 50.0% at six-node difficulty and 0.0% at nine-node difficulty. All four graph interventions produced identical greedy success rates; no task-success benefit from the graph was detected on these episodes.

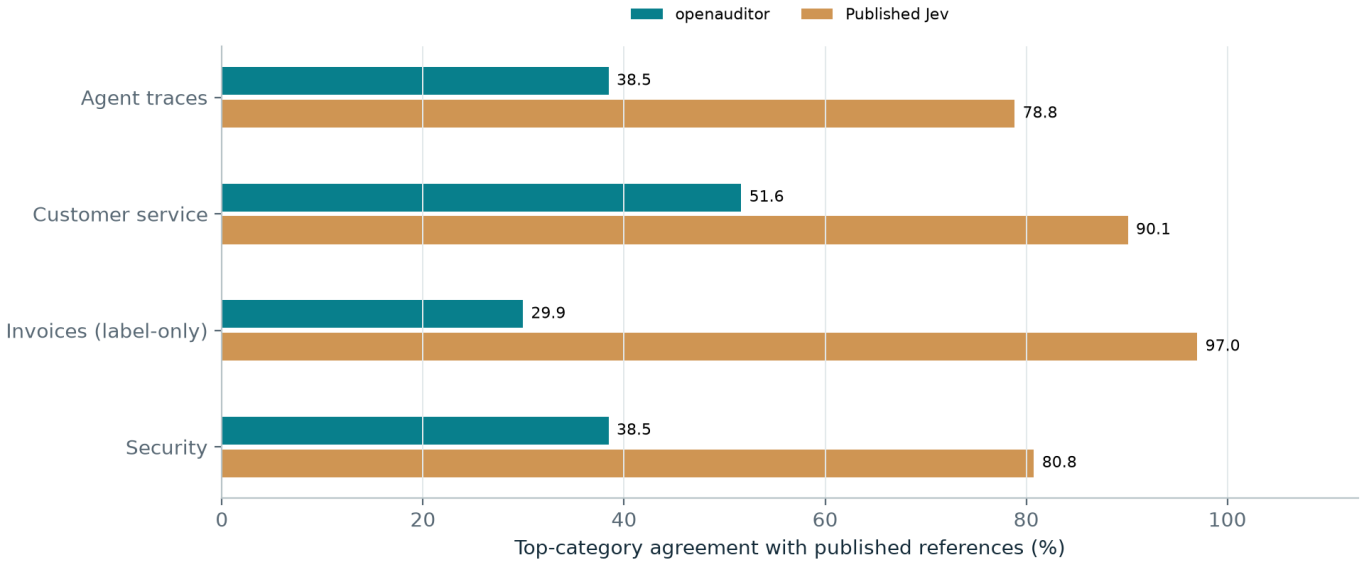
Reinforcement learning

90 updates, 18 with nonzero advantages; maximum sampling/replay log-probability drift 0. Four independently sampled trajectories of the same seeded task supply leave-one-out baselines. The same action temperature is used in sampling and replay. Success forecasts use observed outcomes and proper losses; rehearsal is logged separately.

Ablating a jointly trained graph measures reliance, not superiority over an equally trained graph-free model. Equal results do not demonstrate a GNN benefit. This is an open RL/calibration recipe; TypeSafe's proprietary RLCD algorithm has not been reproduced exactly.

A narrow matched-input check

Five selected cases per workflow domain, not the full published benchmark or a Jev API head-to-head.



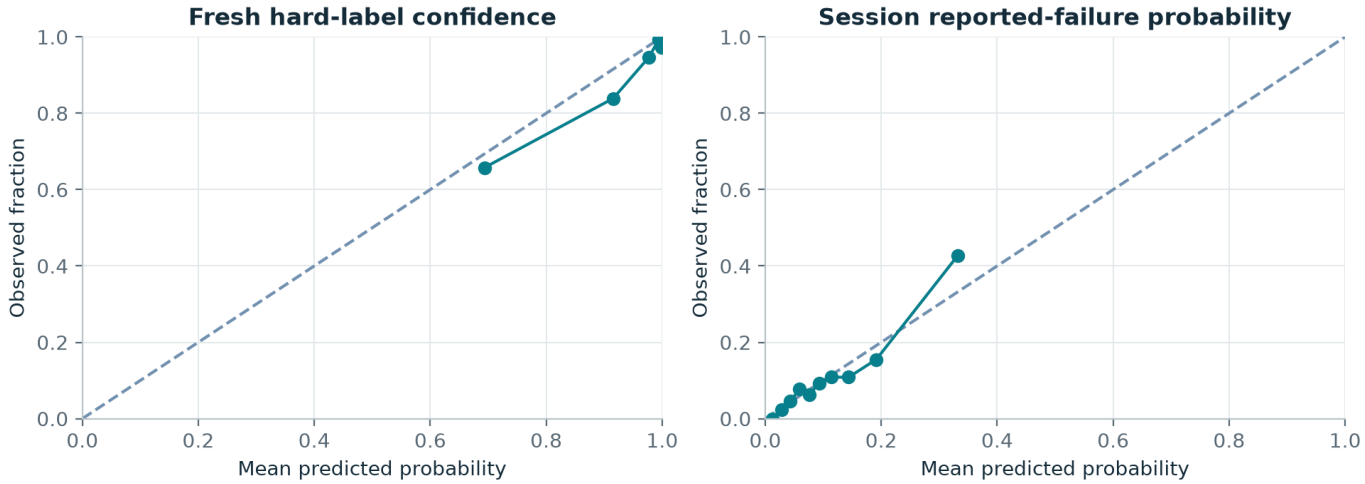
Domain	Comparable questions	Reference
agent trace observability	52	Mean reference distribution
customer service	91	Mean reference distribution
invoice processing	167	Unanimous label-only reference
security incidents	26	Mean reference distribution

The harness replays Jev-visited questions on their published inputs. It does not recompute workflow branches. References are model judgments, not human-verified truth. Seven customer questions lack compatible distributions; invoice references expose labels but no probabilities. Only unanimous, compatible label-only references are scored for invoices. Published probabilities are rounded, and cases were selected by the publisher.

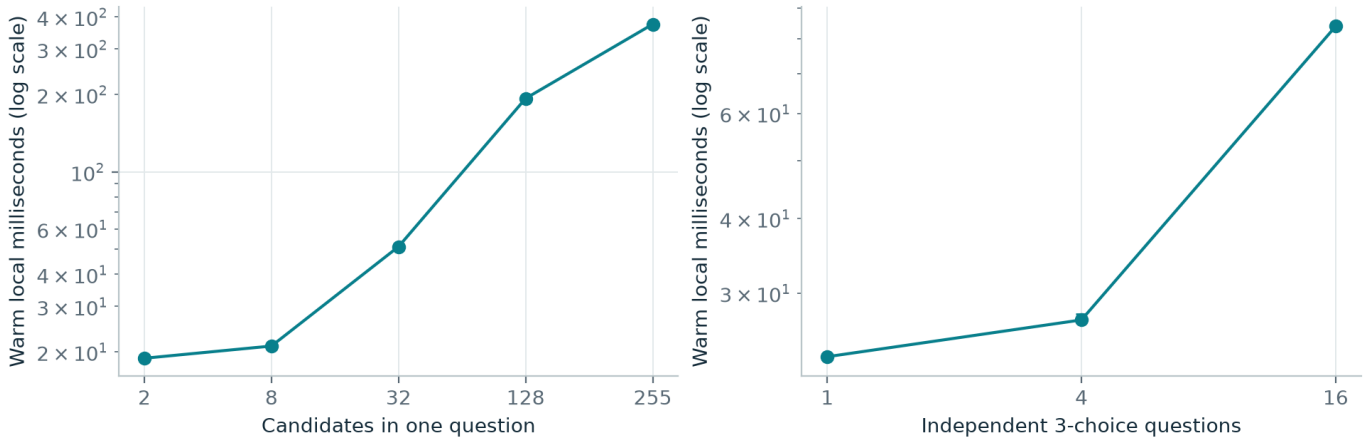
openauditor is measured with its saved configuration, without a runtime context override. No timing comparison is made with the hosted Jev service.

Probabilities and compute cost

Descriptive diagnostics on the frozen model; no test-set calibration or latency matching.



Reliability diagrams use ten equal-count bins. Left: hard-label decisions from the fresh contrast families only. Right: reported execution failure from private-session test cases. Pooled reliability can hide task-specific errors, and paired questions are not independent samples. These are descriptive checks, not confidence intervals or deployment guarantees.



Dots are medians; whiskers span the minimum and maximum of ten warm measurements. This is one H100 with local tokenization, inference and normalization; it excludes model load, network transport and service concurrency. More choices repeat prompt processing. No hosted Jev speed or cost advantage is inferred.

Calibration warning: the fitted Score temperature is 0.05. On 116 test rubric decisions, accuracy is 97.4% but NLL is 3.341: a few mistakes are extremely confident. High accuracy alone does not establish reliable probabilities.

Behavior, speed, reproducibility

One-map game control and bounded link traversal are demonstrations, not universal-agent benchmarks.

Test	Result
Doom: engage (8 episodes)	Kill rate 100.0%
Doom: hold fire (8 episodes)	No-shots-fired compliance 100.0%
Wikipedia navigation	1/6 available routes reached; 0 unavailable

H100 warm local API: 3-choice median **23.5 ms**; 255-choice median **374.3 ms**, 10 measured repeats per case. Tokenization, forward and normalization included; model load, HTTP/network and cold start excluded. Repeated candidate prefill scales with question count; this is not Jev's specialized parallel sampler.

Doom uses visible engine coordinates and explicit control rules, not pixel perception. The new curriculum includes control-shaped tasks. Wiki follows offered links without a teacher, oracle path or revisit override; the same harness is used for both checkpoints. Live link lists are capped and cached for replay. Unavailable external pages are excluded from route-success denominators and shown separately, not counted as model failures.

Reproduce and inspect

55 offline tests passed at delivery. Training configs, optimizer/RNG resume states, dataset hashes, exact stage-source archives, no-text predictions and checksums accompany the private checkpoint. Training continued an initialized adapter with a new optimizer. The export-only context guard was expanded before final tests; learned weights were not changed by that configuration update. All candidate evidence fingerprints must agree.

Canonical model-tree SHA-256:
ec83247177a6919c17944093fdd02d3258906dc809c16b326e0dfd879275efeb

Jev announcement and demo scope
TypeSafe API primitives
Published workflow methodology and cases