ScopeSelect

Risk-Calibrated Context Selection for Healthcare Multi-Agent Handoff

Manu Agrawal and Sumit Mukherjee

Oracle Health

NeurIPS 2026 main track. Poster presentation.

Task handoff creates a memory boundary

Semantic relevance alone cannot decide what another agent should receive.

Prior authorization

A diagnosis may supply evidence for an insurance request.

Appointment scheduling

The same diagnosis may be unnecessary for finding a time.

Authorization determines eligibility.
Task necessity determines what the handoff needs.

Research contribution

Governed memory handoff combines authorization, useful context and evidence of transfer.

Formulation Separate receiver eligibility, unnecessary sensitive exposure and task-failure risk.

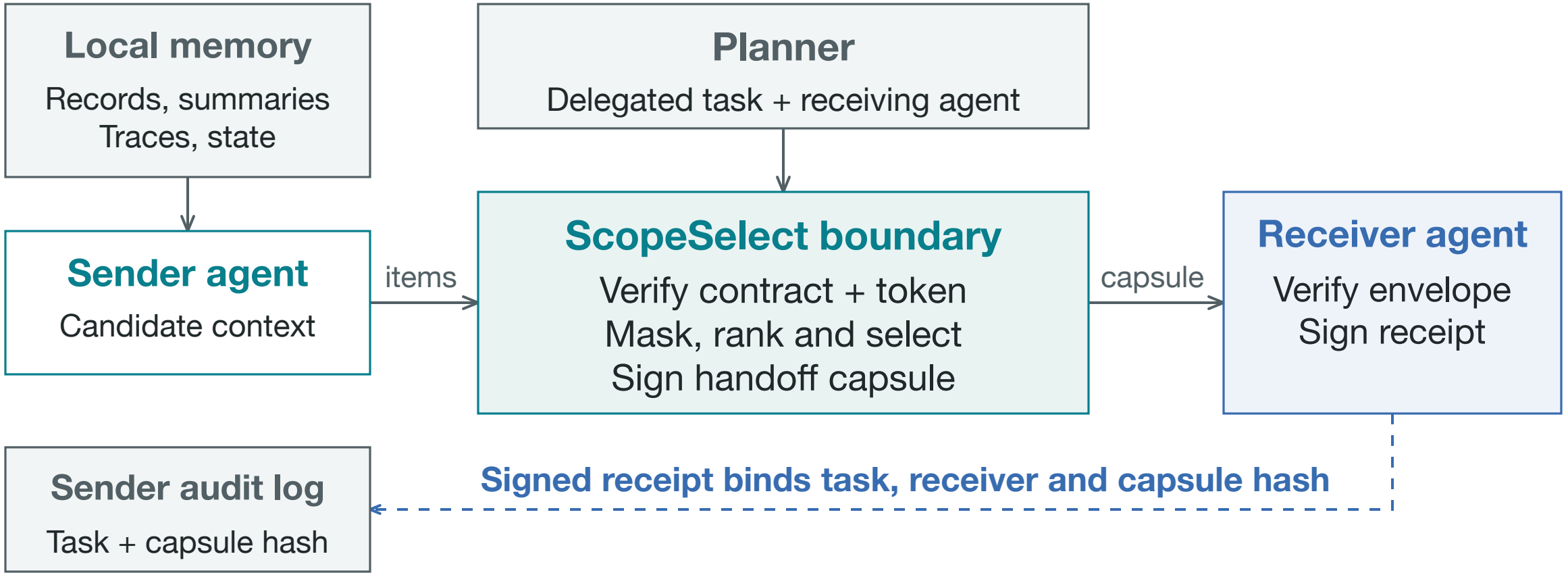
Framework Enforce the boundary around planner proposals, then bind each capsule to signed audit evidence.

Evidence Measure completion and disclosure separately, including failure to transfer calibration to a live receiver.

The contribution applies established optimization and calibration tools to governed handoff.
The experiments evaluate policy simulation, learned selection and cryptographic checks separately.

Architecture: an enforced memory-transfer boundary

Framework protocol. Every handoff and follow-up context request repeats the checks.



Evaluation gap: the top-k harness attempted six infeasible cases instead of escalating.
The cryptographic prototype and selection experiments test different components.

Scheduling example: permission and task necessity

Assigned toy inputs. The contract admits L1 summaries and L2 evidence.

Memory item	Policy	Role in toy task	Score
Appointment intent	Allowed L1	Required, nonsensitive	0.95
Preferred time	Allowed L1	Required, nonsensitive	0.90
Contact channel	Allowed L1	Required, sensitive	0.85
Referral summary	Allowed L2	Unneeded, sensitive	0.30
Restricted notes	Blocked L3	Outside authorized set	--
Private agent reasoning	Blocked L4	Outside authorized set	--

Only the first four items enter ranking. The mask excludes restricted notes and private state regardless of a relevance model's score.

Toy handoff: completion versus unnecessary exposure

Required fields: appointment intent, preferred time and contact channel.

Authorized capsule	Coverage	Task outcome	Exposure (items)
Top 2: intent + time	2 / 3	Fail: no contact	0
Top 3: add contact	**3 / 3**	**Pass**	**0**
All 4: add referral	3 / 3	Pass	1

Contact is sensitive but necessary, so it adds no unnecessary exposure.
The extra referral summary adds one unnecessary sensitive item.



Illustrative outcomes, not measured results. Escalate if the task needs context the contract excludes.

Learning and calibrating the context threshold

k is the number of ranked authorized items retained. Freeze the ranker before calibration.

Learn a task-item ranking

Train the logistic scorer on item-necessity labels.
Features include required-field match.

Conformal Risk Control (CRC)

Choose the smallest k with B(k) ≤ α.
α targets expected task-failure loss.

B(k) = (failures(k) + 1) / (n + 1)

None qualifies: retain an escalation policy.

99 assigned toy calibration cases

k	Failures	B(k)
1	45	0.46
2	19	0.20
3	**9**	**0.10**
4	0	0.01

Target α = 0.15. Toy choice: k = 3.

CRC requires exchangeability, per-case nonincreasing loss and a maximum rule with loss ≤ α.
Rejections counted as failures can break this endpoint condition. The statistic alone is not a certificate.

Evaluation settings and the task-success rubric

Policy simulation

21,600 tasks

4 workflows across 6 seeds
Stochastic completion simulator

Learned selection

3,600 cases

2,160 train / 720 calibration / 720 test
Deterministic field-coverage scorer

Deterministic success = valid verification + every required field present.

Escalations count as failures. Exposure counts unnecessary sensitive items.

Synthea-HH adds 340 generated FHIR bundles and 8,609 memory items for structure and label audits. These settings provide no clinical-outcome or end-to-end deployment validation.

Held-out selection: success and exposure

720 held-out cases. Every method shares the hard authorization mask.

Method	Success	Exposure*	Tokens
Authorized retrieval, all	94.9%	6.08	4,995
ScopeSelect, k = 8 (cal.)	**94.7%**	**1.81**	**2,248**
LLM reranker, k = 4 (cal.)	87.6%	0.43	1,236
LLM reranker, k = 6 (grid)	94.9%	1.25	2,133
LLM selector, uncalibrated	64.9%	0.38	1,181

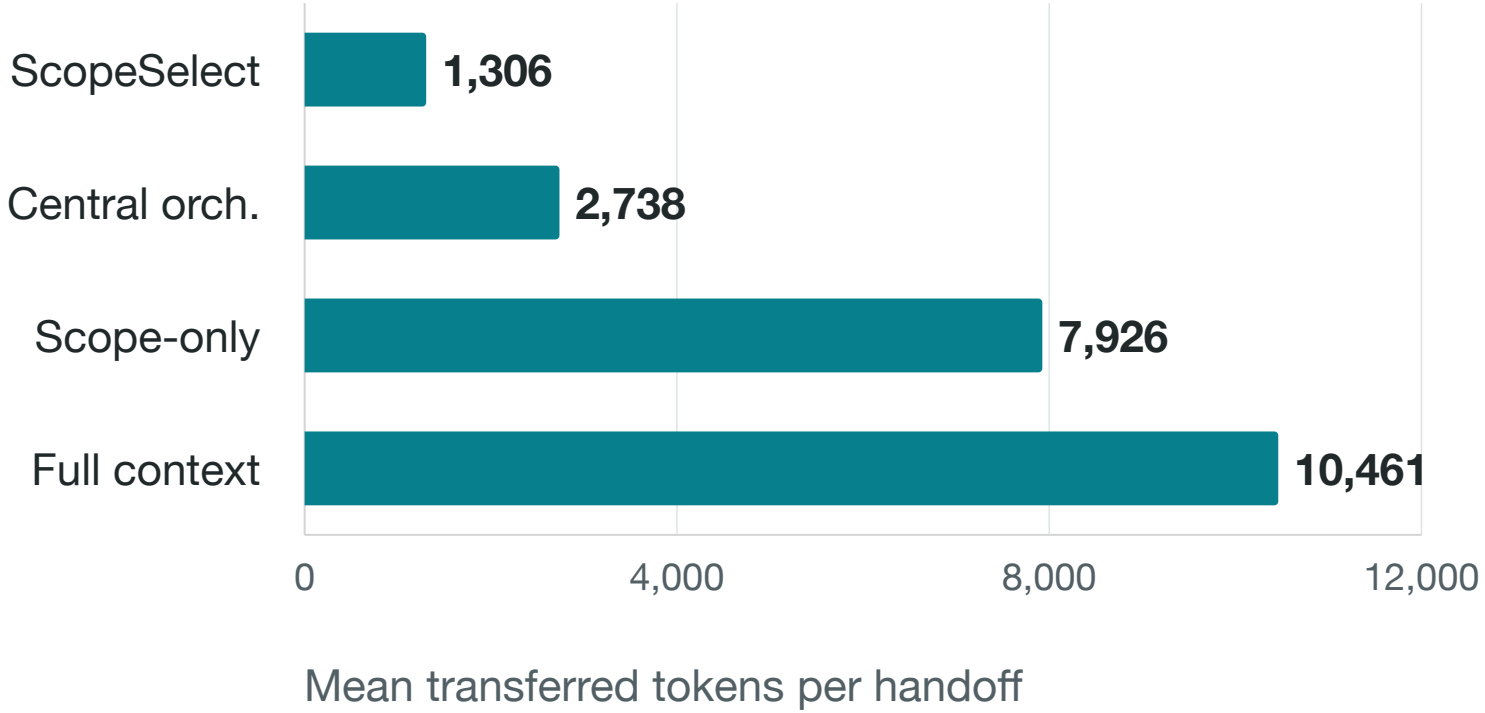
* Mean unnecessary sensitive items per handoff. Cal. = calibration-selected.

Vs authorized-all: exposure falls 70.3% and tokens 55.0%. Success is 94.7% vs 94.9%.

The reranker has a stronger k = 6 grid point. The framework can govern either authorized selector.

Simulation measures the policy tradeoff

Six-seed means. Simulated completion is separate from the learned-selection experiment.



79.0%

autonomous success
83.8% for full context

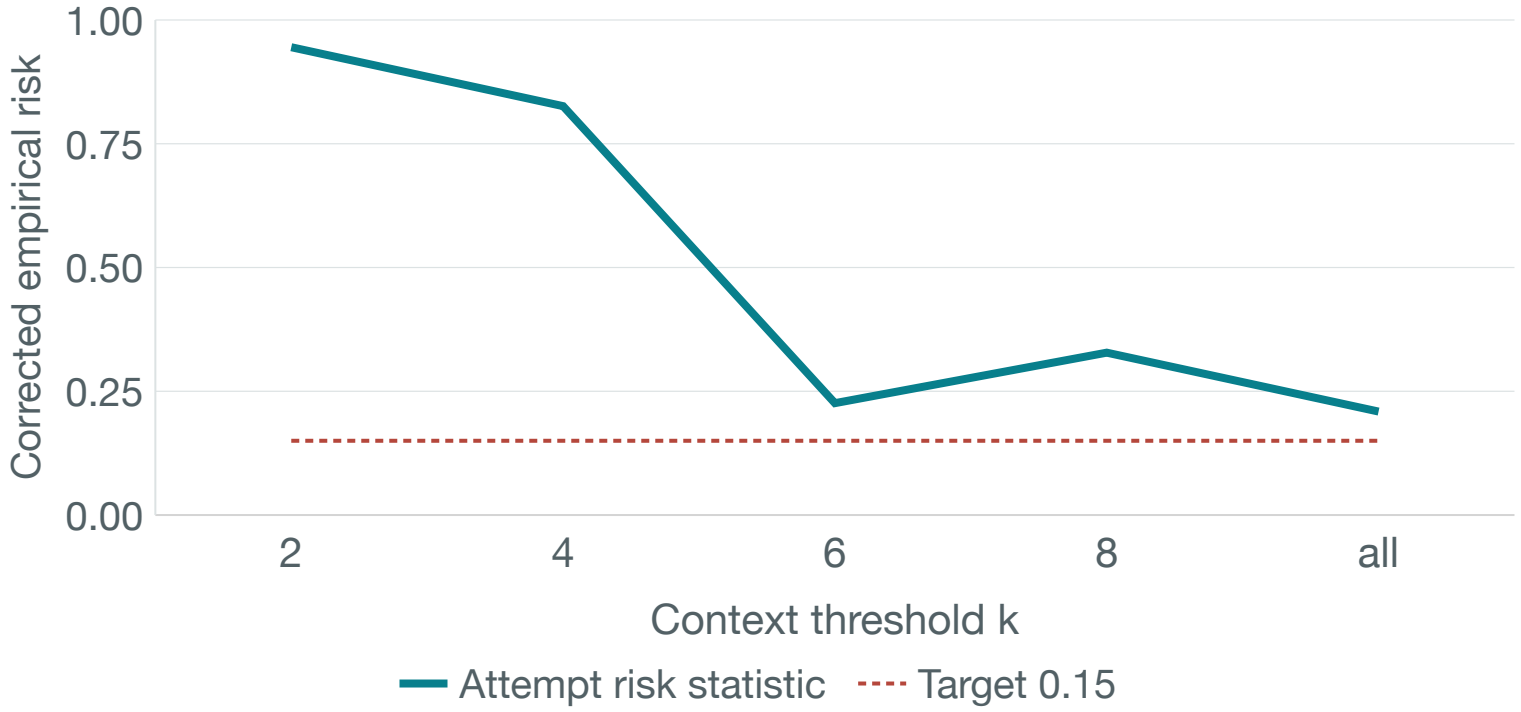
7.2%

escalation rate

87.5% fewer tokens than full context. The 0.00 exposure row uses clean policy labels and deterministic filtering, rather than learned zero-exposure generalization.

Live receivers need separate risk validation

250 calibration + 250 test cases. Curve: 234 calibration attempts, excluding 16 escalations.



0.259

lowest all-case calibration statistic (includes escalations)

Attempts-only minimum: 0.209
Non-monotone loss

Both populations exceed the 0.15 target. Observed non-monotone loss prevents applying the monotone CRC theorem to this live selector-receiver family.

Trust boundary and security scope

Transfer-time enforcement depends on correct policy inputs and a sound verifier.

Protocol checks

Identity, audience, scope and freshness
Replay and delegation-token validity
Receiver-permitted memory layers

Required trust

Policy tags and contract semantics
Signing keys and the verifier
Enforcement on each context request

Outside the established guarantee

Compromised policy services, collusion, covert channels and downstream redisclosure.

Policy-tag stress (4,320 cases, $p = 0.30$): under-restrictive layers yield 9.07% unauthorized release.
Injected corruption is a stress model, not a measured classifier-error rate.

Executed protocol checks and local prototype

50 / 50

Tier-A defense assertions pass

10 handcrafted attack families
50 Tier-B scenarios add coverage
of 27 further families

2,000

local signed-handoff trials

1,434 valid accepted
566 unsafe rejected
0.79 ms median valid verification

Tier B is coverage-only, without an executed defense rate. Local timings exclude network, model calls and production key management.

An explicit, testable memory-transfer boundary

ScopeSelect links receiver authorization, selected context and signed handoff evidence.

Held-out selection reduces exposure 70.3% and tokens 55.0% versus authorized full context. Strong rerankers remain competitive.

Deployment requires reviewed policy inputs and validation of the selector and receiver together.

https://openreview.net/forum?id=FLlWDrOo4c

agrawalmanu20@gmail.com

Backup: budgeted selection and top-k evaluation

U(S) = weighted field coverage + learned item rewards

Assumptions: normalized, nonnegative, monotone submodular U.
Nonnegative additive costs, with $D(S) \leq B$.

Modified cost-benefit greedy over authorized items,
compared with the best feasible singleton, has a
classical $\frac{1}{2}(1 - 1/e)$ objective-value bound.

Classical result: Khuller et al. (1999), Leskovec et al. (2007).

The calibration results evaluate a separate learned top-k implementation.
The objective bound does not imply top-k task success or a live-receiver risk guarantee.

Backup: metric definitions

Metric	Definition
Autonomous success	Completed delegated task under the stated rubric
Escalation	Case referred for non-autonomous handling
Unnecessary sensitive exposure	Mean sensitive items not needed for the task
Unauthorized release	Transferred item outside reference authorization
Audit completeness	Required audit fields present
Corrected CRC statistic	(Calibration failures + 1) / (Calibration cases + 1)

Backup: independent label audits

Small, separate rubric audits. Agreement varies substantially across label types.

Audit sample	Necessity κ	Sensitivity κ	Authorization κ
Benchmark, 50 items	0.63	0.80	--
Synthea-HH, 50 items	–0.18	–0.14	--
Expert check, 88 rows	0.243	0.856	–0.020

Sensitivity agreement is strongest in the expert check.
Necessity and authorization expose contested policy boundaries.

-- = not shown for that audit. These checks assess the rubric, not clinical performance.

Backup: unified held-out metrics

720 test cases. Reranker k = 6 is an extra grid point. The direct LLM selector is uncalibrated.

Method	Success	Esc.	Exposure	Tokens	Unauth.	Audit flag
Authorized all	94.9%	4.3%	6.08	4,995	0%	100%*
ScopeSelect k = 8	**94.7%**	**4.3%**	**1.81**	**2,248**	**0%**	**100%***
Reranker k = 4	87.6%	4.3%	0.43	1,236	0%	NR
Reranker k = 6 (grid)	94.9%	4.3%	1.25	2,133	0%†	NR
LLM selector, uncal.	64.9%	4.3%	0.38	1,181	0%†	NR

* Harness flag set by construction, not executed receipt verification. NR = not recorded.
† Shared-mask property; no independent selected-ID replay is claimed here.

Calibration statistic: ScopeSelect 0.064; selected reranker 0.126. These are not test risks.

Backup: case-matched escalation

21,600 simulation cases, including 1,558 observed escalations.

Workflow	Esc. rate	Verification	Insufficient	Sufficient subset
Prior authorization	6.57%	338	3	14
Discharge coordination	7.65%	373	40	0
Patient support	8.35%	368	83	0
Health-IT incident	6.28%	339	0	0
All workflows	7.21%	1,418	126	14

14 / 1,558 referrals (0.90%) had sufficient authorized context under the simulator's 86% coverage rule; the ScopeSelect policy requires every necessary item.

Backup: policy-metadata errors

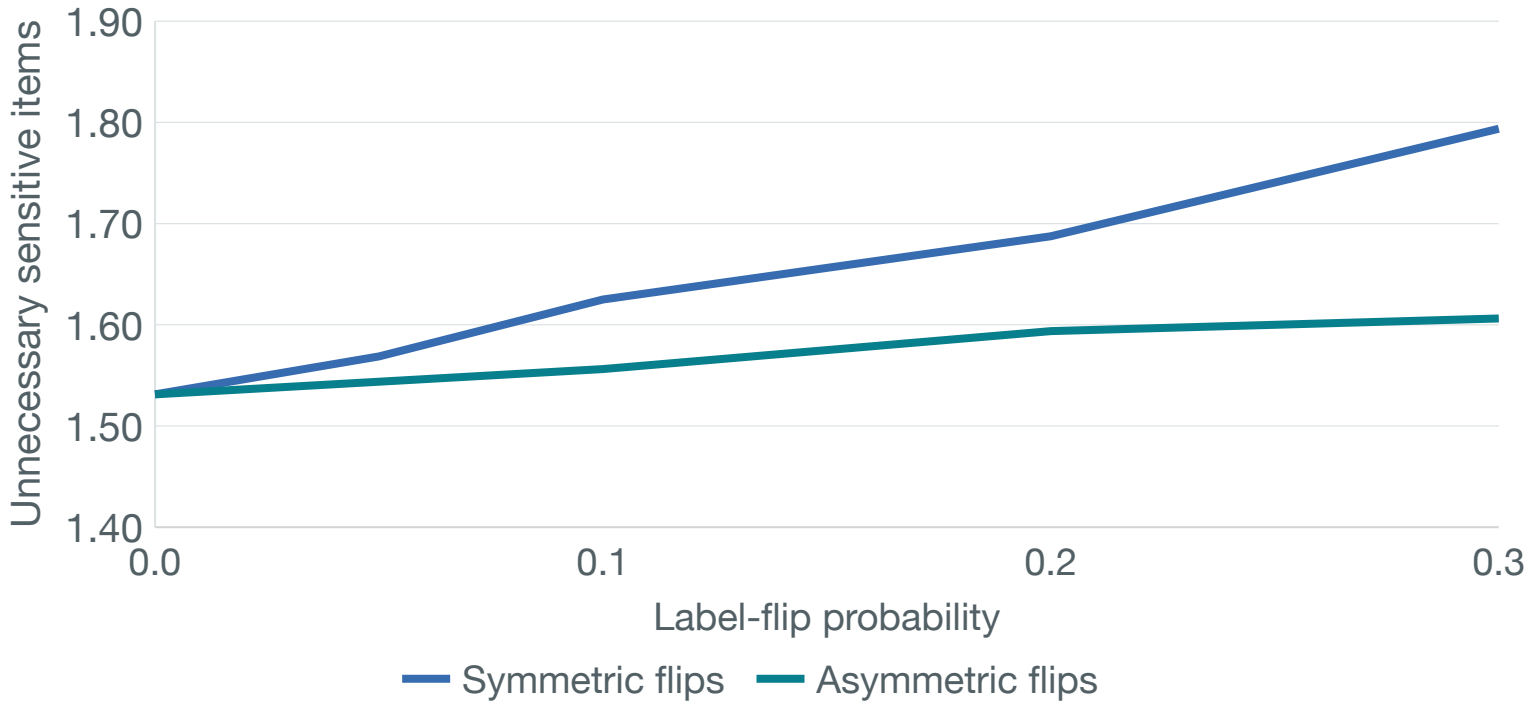
Six seeds; 4,320 test cases; fixed k = 8; injected corruption probability 0.30.

Observed metadata	Success	Unauth. release	Exposure
Clean metadata	92.48%	0.00%	1.765
Under-restrictive layers	92.57%	9.07%	1.764
Over-restrictive layers	39.63%	0.00%	3.100
Receiver L3 permission	92.59%	6.92%	1.764

Outcomes use clean reference authorization. Injected errors quantify metadata dependence; they do not estimate real classifier accuracy.

Backup: selector-feature noise

Single-seed, fixed-k stress test. Structural authorization labels remain unchanged.



89.4%

deterministic success
at every tested rate

These perturbations change necessity/sensitivity features, not the authorization boundary.
Policy-tag failures are measured separately on the preceding slide.