

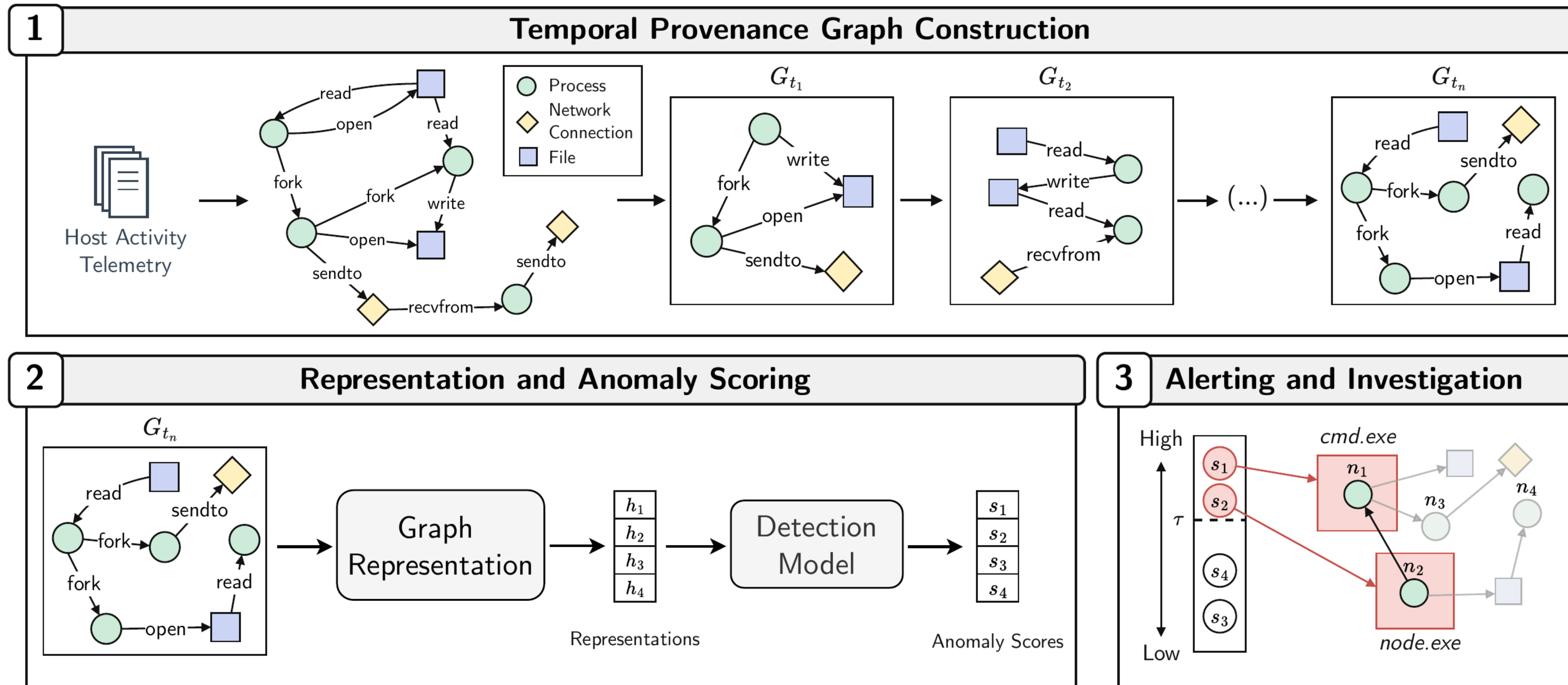
How Benchmarks and Evaluation Protocols Shape Conclusions in Provenance-Based Intrusion Detection

Lorenzo Guerra, Thomas Chapuis, Guillaume Duc, Pavlo Mozharovskyi, Van-Tam Nguyen, NDSS 2027



Why provenance graphs?

- Expressive data, useful for training neural networks and interpreting malicious behavior
- Causal view of execution: processes, files and network flows linked by system events.



PIDS report strong results, but conclusions depend on how they're measured

Benchmark	Protocol	Metrics
Does the dataset carry enough signal to compare models?	Are checkpoints and thresholds chosen without the test set?	Do they measure alerting, or investigation?

We test which architectural claims survive one shared protocol.

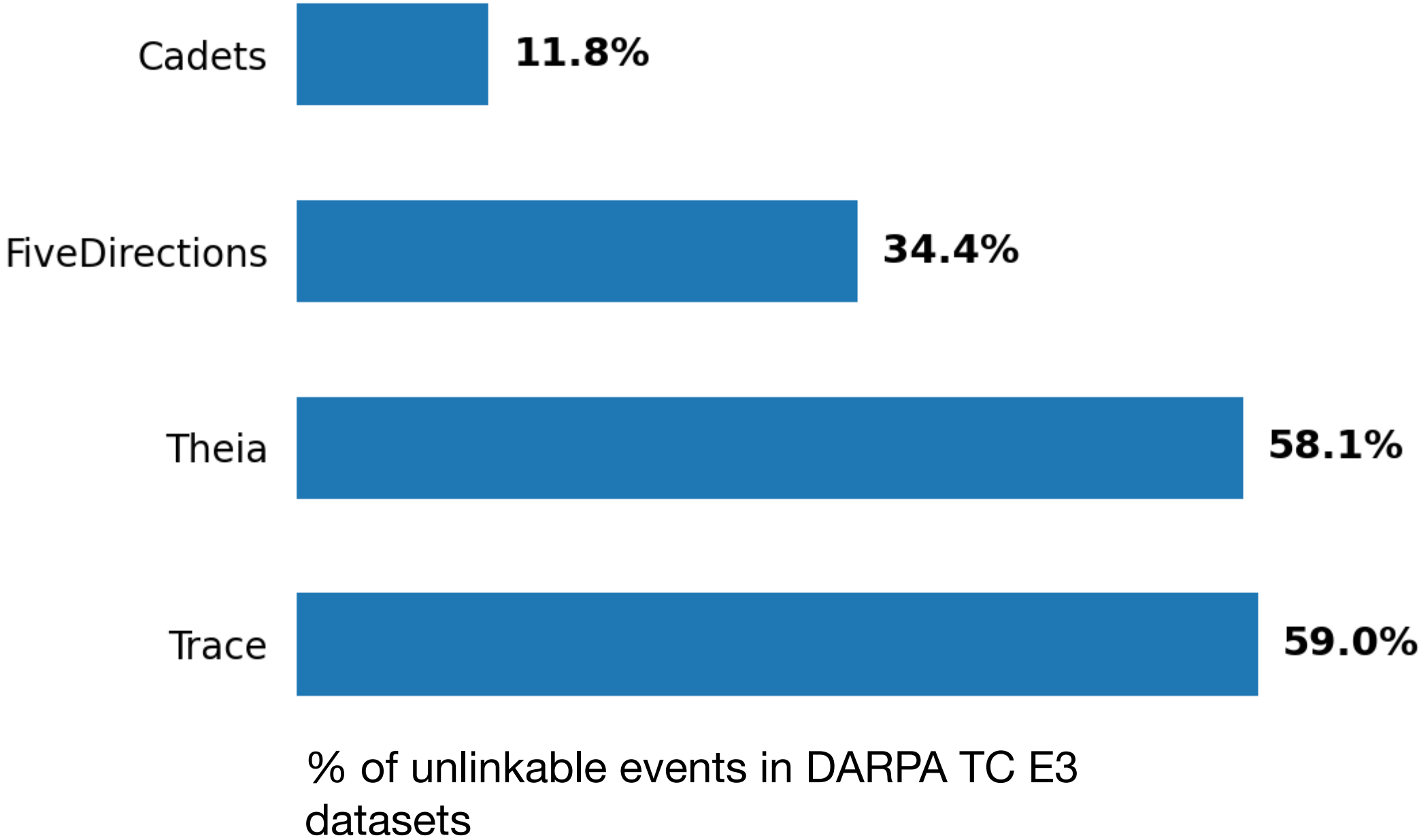
Up to ~59% of E3 events can't be linked into the graph

Benchmark

Protocol

Metrics

Missing object UUIDs and unresolved entity references cause provenance relationships to be lost.



Only audited E3 meets all three evaluation requirements

Benchmark	Protocol	Metrics
-----------	----------	---------

- Reproducible process-level labels
- Public artifact audit
- Validation set containing independent attacks

Benchmark	Compatible labels	Dataset audit support	Attack validation	Role
Audited E3	Yes	Yes	Yes	Primary
ATLASv2	Yes	Yes	No	Secondary
E5	No	No	Varies	Excluded
OpTC	No	Yes	No	Excluded

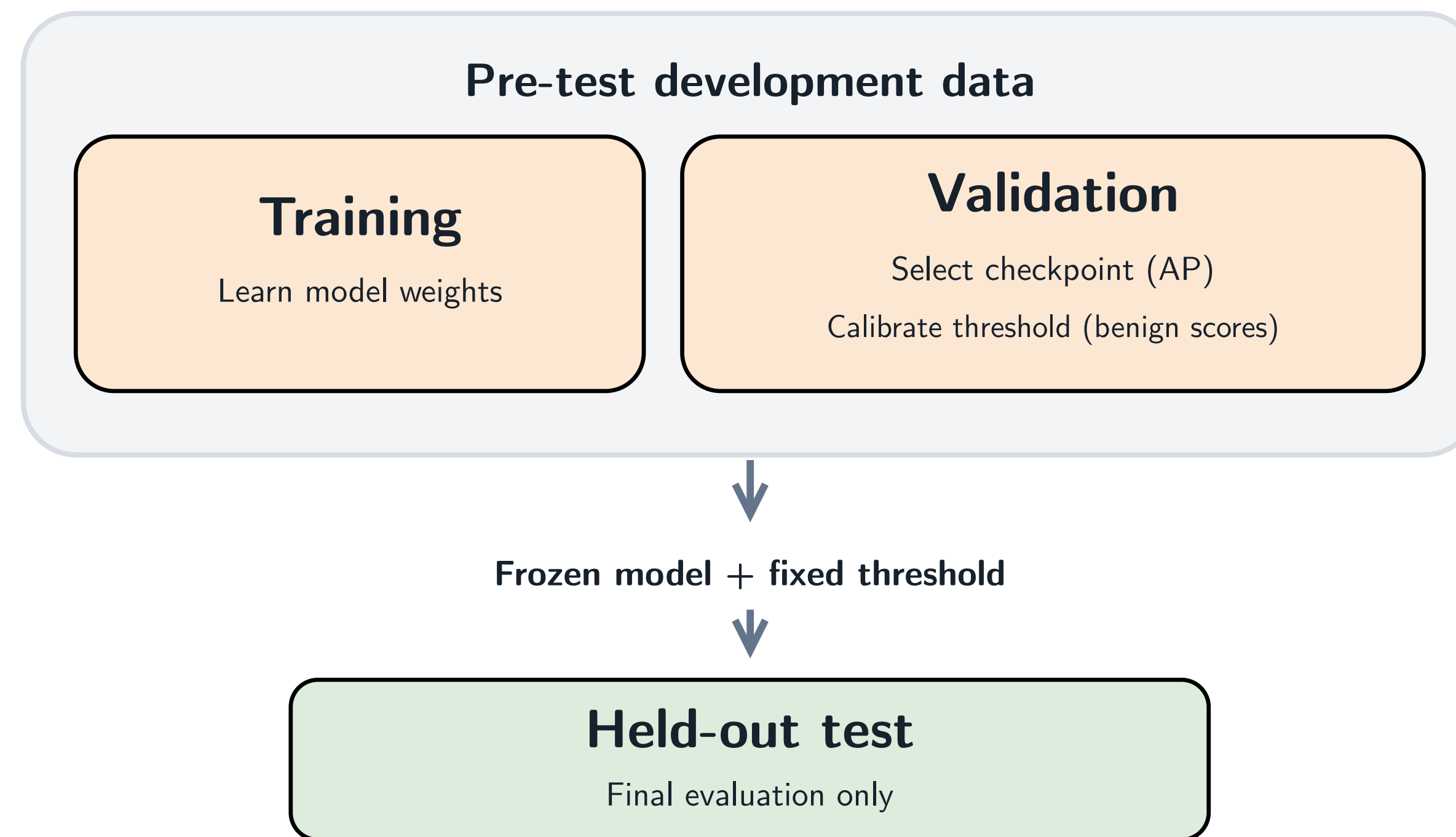
Model selection and calibration never use the test period

Benchmark

Protocol

Metrics

- Checkpoint: selected using validation AP
- Threshold: max benign validation score



High precision does not imply useful investigation

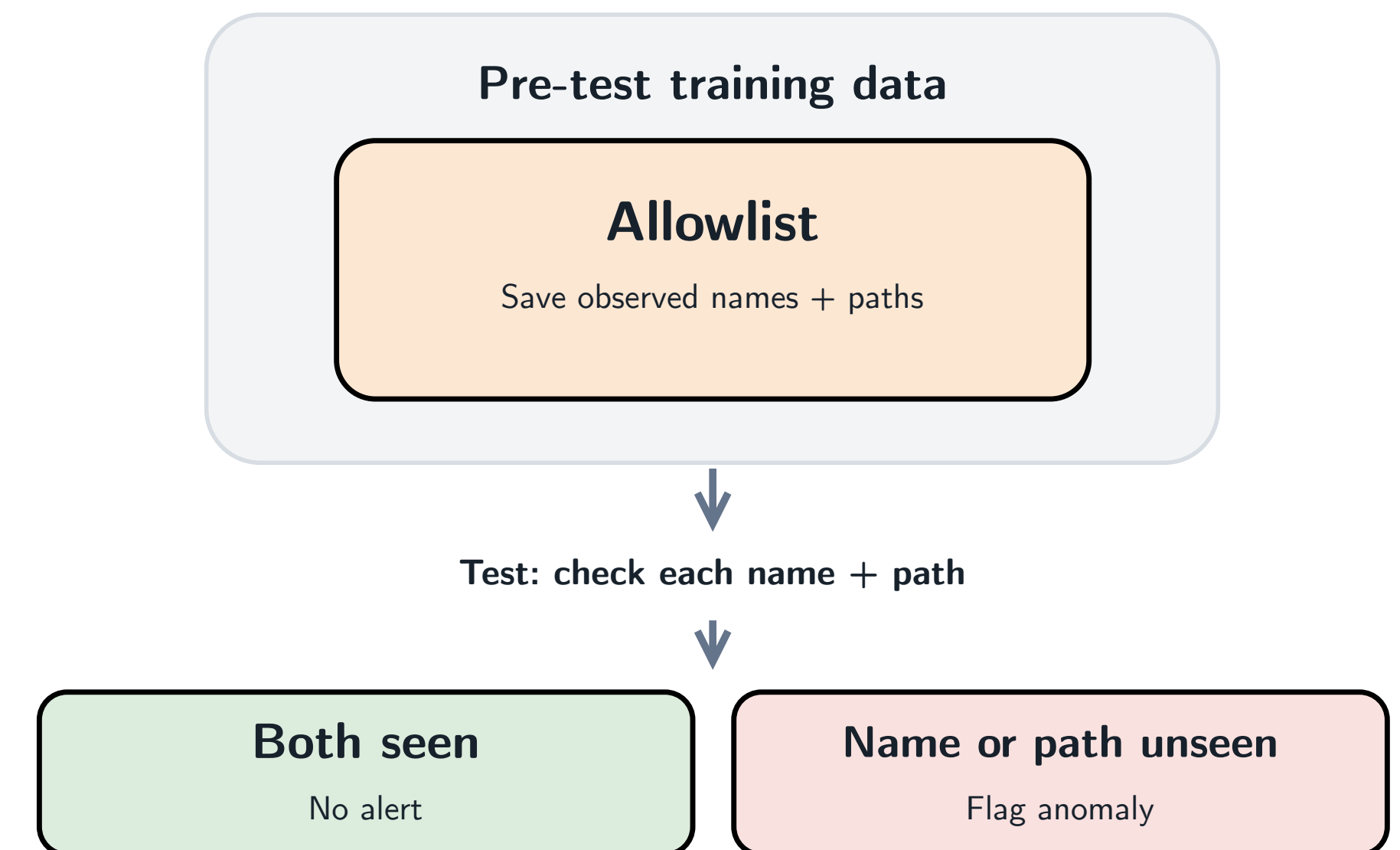
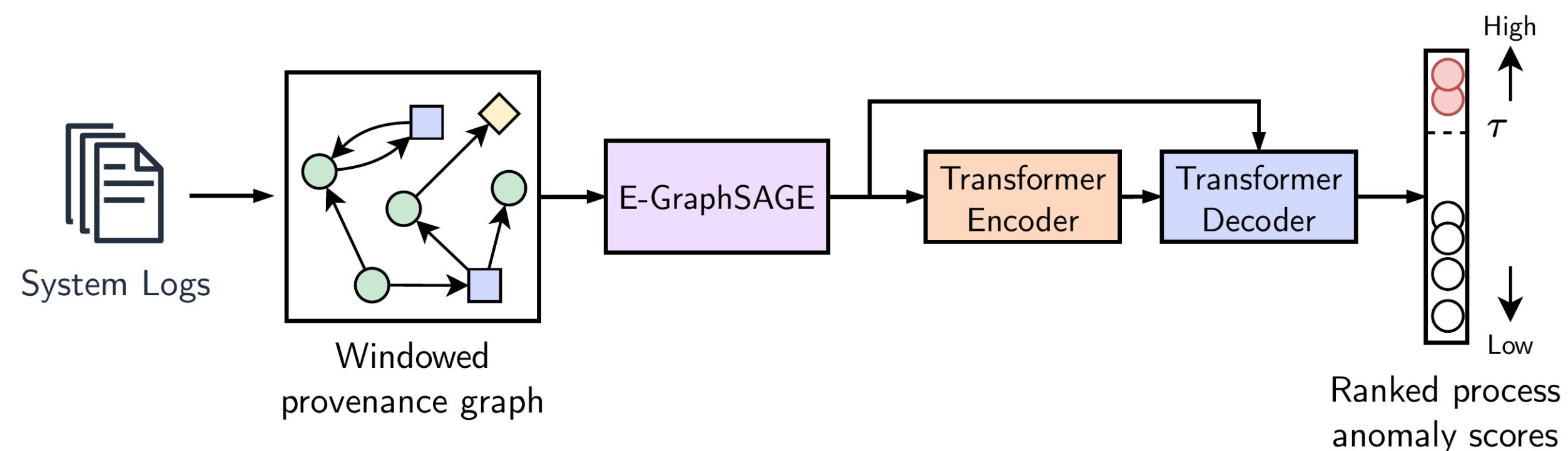
Benchmark	Protocol	Metrics
-----------	----------	---------

- A detector can raise accurate alerts while recovering almost none of the attack footprint.
- Precision and ADP measure alerting; F1 and recall measure investigation.
- Threshold-independent metrics (AP, AUROC) measure ranking quality.

Dataset	Model	F1-score	Precision
Cadets E3	Velox	0.007 $\pm$ 0.004	0.883 $\pm$ 0.162
	Orthrus	0.005 $\pm$ 0.006	0.600 $\pm$ 0.548

What do learned graph models add beyond lexical novelty?

- Theseus adapts joint graph reconstruction to windowed provenance graphs.
- Compare it with Orthrus, Velox, Magic, and an executable/path allowlist.
- Allowlist: flag a process if its executable name or path was never seen during training.



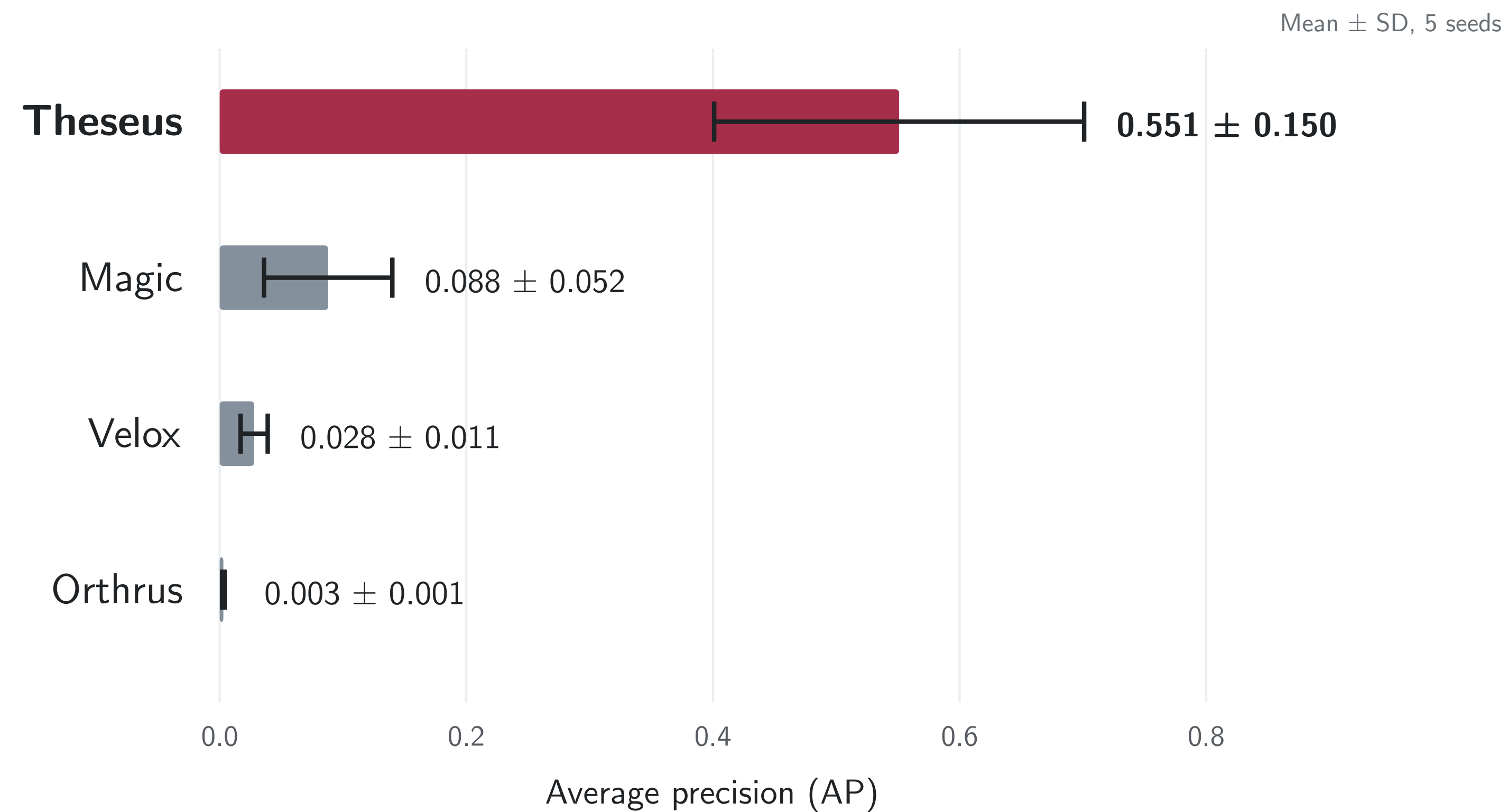
A simple allowlist wins at the conservative operating point

The allowlist achieves the highest F1 and MCC on all four audited E3 datasets.

Dataset (<i>Q</i>)	System	F1 ↑	MCC ↑
Cadets (1.74)	Allowlist	0.021 ± 0.000	0.058 ± 0.000
	Orthrus	0.005 ± 0.004	0.036 ± 0.033
	Velox	0.007 ± 0.006	0.048 ± 0.022
	Magic	0.000 ± 0.000	0.000 ± 0.000
	Theseus	0.000 ± 0.000	-0.001 ± 0.002
FiveDir. (3.97)	Allowlist	0.006 ± 0.000	0.011 ± 0.000
	Orthrus	0.000 ± 0.000	0.000 ± 0.000
	Velox	0.000 ± 0.000	0.000 ± 0.000
	Magic	0.000 ± 0.000	0.000 ± 0.000
	Theseus	0.000 ± 0.000	0.000 ± 0.000
Trace (2.95)	Allowlist	0.107 ± 0.000	0.216 ± 0.000
	Orthrus	0.000 ± 0.000	0.000 ± 0.000
	Velox	0.000 ± 0.000	0.000 ± 0.001
	Magic	0.000 ± 0.000	0.000 ± 0.000
	Theseus	0.008 ± 0.017	0.016 ± 0.037
Theia (6.72)	Allowlist	0.273 ± 0.000	0.324 ± 0.000
	Orthrus	0.000 ± 0.000	-0.001 ± 0.000
	Velox	0.097 ± 0.001	0.102 ± 0.002
	Magic	0.000 ± 0.000	0.000 ± 0.000
	Theseus	0.251 ± 0.186	0.306 ± 0.198

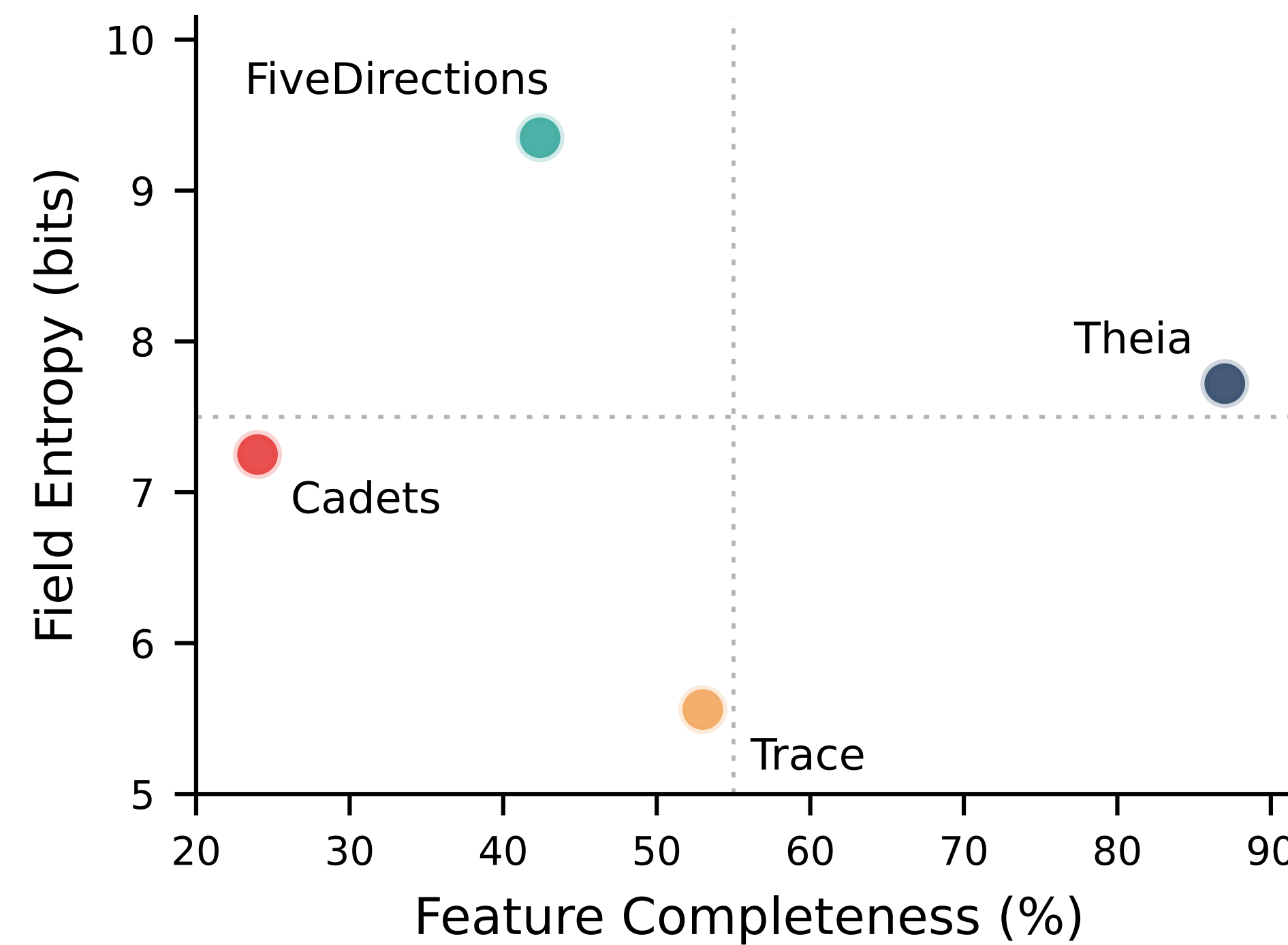
Theia is the exception: richer modeling improves detection

Theseus achieves 0.551 AP, versus 0.088 for the strongest competing learned baseline.



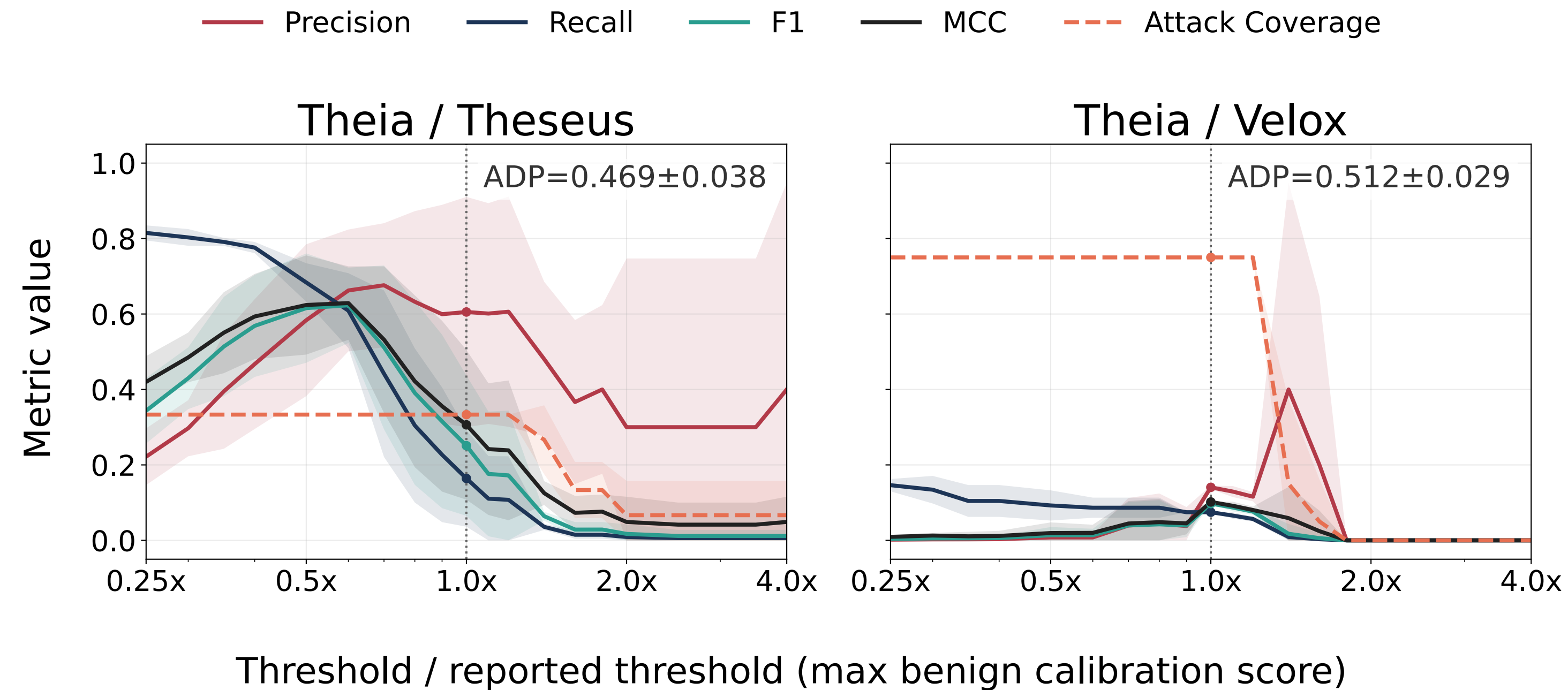
Theia provides richer semantic information

87% feature completeness, with the strongest combination of populated attributes and semantic diversity.



The conservative threshold hides recovery potential

A post-hoc sensitivity analysis on Theia shows that reducing Theseus's threshold to 0.6x increases F1 from 0.251 to 0.623.



How to evaluate PIDS claims

- Check the quality of the data before comparing architectures.
- Calibrate and select checkpoints on validation only.
- Report both alerting and investigation metrics.
- Always include a simple baseline (e.g. the allowlist).

Better benchmarks are needed to design better models

- **Graph fidelity:** resolved dependencies and complete semantic attributes
- **Higher-quality features:** permissions, privileges, return value
- **Realistic workloads:** documented shifts and independent validation
- **Precise ground truth:** granular labels with explicit ambiguity

Thank you