Session ID: PLA1601

Log Reasoning Model: Grounded AI for Logs

.conf26 breakout | Log Pruner and AI Detection Agent use cases

Nick Ma
Senior Staff Product Manager, Splunk Central AI

Leo Betthauser
Senior Staff Machine Learning Engineer, Splunk Central AI

SESSION FLOW

Agenda

Use cases first; technical depth second.

01	**Why log reasoning matters**	The evidence bottleneck between detection/search and AI reasoning.
02	**Log Pruner for O11y RCA**	Reduce noisy log volume into RCA-ready context inside the LLM window.
03	**AI Detection Agent for unknown threats**	Turn weak risk signals into reviewable findings with confidence and provenance.
04	**Technical deep dive**	How the encoder, training data, task heads, and service architecture work.

© 2026 Cisco and/or its affiliates. All rights reserved.

Cisco Confidential

Source: pitch + CiscoLive

Log reasoning turns noisy machine data into grounded evidence for AI decisions.

It is the cognitive process experts use to extract meaning and make decisions from machine-generated log data. The model makes that evidence-selection layer reusable.

© 2026 Cisco and/or its affiliates. All rights reserved.

Cisco Confidential

Source: pitch deck slide 26

The investigation workflow has not changed

The expensive part is still finding the right evidence before reasoning begins.



Where time disappears

- Manual query pivots
- Repeated filtering
- Cross-silo context hunting

Where cost disappears

- Too many logs in the context window
- Too many LLM retries
- Redundant evidence

Where trust disappears

- Missing provenance
- Unclear confidence
- No reviewable chain

© 2026 Cisco and/or its affiliates. All rights reserved.

Cisco Confidential

Source: pitch deck slides 19, 27

Why frontier LLMs alone do not solve logs

The LLM should reason after a log-native harness has selected and grounded evidence.

Context overflow

High-volume incidents can exceed the model window.

Blunt truncation silently drops tail evidence.

Signal burial

Relevant clues are sparse among duplicates and noise.

Generic ranking misses log structure and operations semantics.

Brittle translation

Customer topology, sourcetypes, and SPL context matter.

Prompting alone cannot learn tenant-specific evidence habits.

0.6%
relevant-log density in one production case

43K
raw logs before pruning

1K
logs kept by blunt truncation

<1%
target request failure rate

© 2026 Cisco and/or its affiliates. All rights reserved.

Cisco Confidential

Use cases:
from evidence selection to reviewable findings

Source: pitch slides 30, 35

Use case map: where log reasoning lands

One evidence layer can support multiple AI product surfaces.

Log Pruner in O11y

Trigger: incident RCA is overwhelmed by noisy logs.

Output: RCA-ready context that fits inside the LLM window.

Proof: matched quality with 51% lower latency.

AI Detection Agent

Trigger: weak or intermediate ES risk events.

Output: verdict, confidence, evidence chain, MITRE context.

Proof: 30%-40% true positives for unknown candidates in PoC.

Reusable primitives

Semantic log search

Representative subsampling

Key-event summarization

Next-event prediction

If the workflow is 'find the right logs, then reason,' it is a candidate Log Reasoning workflow.

© 2026 Cisco and/or its affiliates. All rights reserved.

Cisco Confidential

Source: pitch slides 15-23

Use case 1: Log Pruner

Large log volume -> RCA-ready context within the LLM window.

Job to be done

As an O11y AI engineer or RCA agent builder, reduce the log set to the most relevant evidence.

Keep the root cause while avoiding context overflow and repeated LLM calls.

Current failure modes

Context exceeded

Noisy duplicates

Manual triage before AI can reason

Blunt truncation hides tail evidence

Product result

Initial model shipped internally to O11y SRE Agent.

Offline and online evals gate rollout.

Same primitive expands into direct search verbs.

93%
fewer logs in initial result framing

51%
lower latency vs baseline

0.737
avg score with LogPruner+LRFM

25-100
high-signal logs passed to reasoning

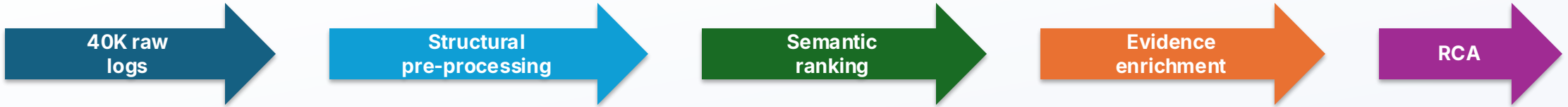
© 2026 Cisco and/or its affiliates. All rights reserved.

Cisco Confidential

Source: pitch slide 21

Prune first, reason second

Compress the haystack before asking the LLM to explain it.



40K raw logs -> about 500 structured representatives -> 25-100 high-signal logs -> grounded RCA

LogPruner handles structure. LRFM handles relevance. Evidence enrichment adds counts and field summaries.

Blocks AI Assistant?

No. The path has graceful fallback.

Integration point

Right before LLM reasoning.

Operating range

25-100 logs passed onward.

© 2026 Cisco and/or its affiliates. All rights reserved.

Cisco Confidential

Source: pitch slide 22

Log Pruner results: matched quality, half the latency

Early evals show the model reduces cost and still holds up on harder cases.

Method	Avg score	Latency	Pruning LLM calls
Baseline	0.747	50.0s	5+
LogPruner + LRFM	0.737	24.7s	0

51%

faster than baseline

0

LLM calls for pruning

Harder 30-case set

Baseline avg score: 0.626

LogPruner + LRFM avg score: 0.661

Only method that held quality as difficulty scaled

What the room should remember

The system does not trade accuracy for speed.

It removes the evidence-selection burden from the LLM.

It lowers latency before the agent reasons.

© 2026 Cisco and/or its affiliates. All rights reserved.

Cisco Confidential

Source: pitch slide 14 + AI deck

Use case 2: AI Detection Agent

Risk-event triggered detection -> reviewable finding for unknowns.

Job to be done

When a low-confidence risk event fires on an entity, pull context and stitch the evidence chain.

Return a verdict with confidence so the analyst can approve, reject, or file the finding.

Why it matters

Surfaces high-confidence unknown findings.

Reduces manual hunting after intermediate risk signals.

Explains MITRE and kill-chain context.

PoC outcome

30%-40% true positives for unknown candidates with SGS validation.

Aligned with ES direction.

Preview path targeted for ES 8.8.

Not replacing rules. Finding the sequences rules miss or fail to escalate.

© 2026 Cisco and/or its affiliates. All rights reserved.

Cisco Confidential

Source: AI deck slides 5, 14

How AI Detection works

A customer-facing technical view of the offline preview run.



Log reasoning contribution

- Select high-signal evidence
- Summarize event-level meaning
- Preserve source provenance for review

Task-specific contribution

- Recognize MITRE-consistent behaviors
- Identify key events
- Separate support from refutation

Graph reasoning contribution

- Connect entities over time
- Score suspicious paths
- Package confidence and reason codes

© 2026 Cisco and/or its affiliates. All rights reserved.

Cisco Confidential

Source: AI deck slides 7-9

Example finding patterns

Fake/scrubbed examples from the PoC show how weak signals become a reviewable chain.

Credential exposure -> remote access -> first-time sudo

Before: three unrelated medium/noisy alerts.

After: plausible intrusion path across the same entity over time.

MITRE context: credential access, remote tool, privilege escalation.

IAM enumeration -> access-key persistence

Before: low-severity IAM fragments.

After: reconnaissance followed by persistence.

The sequence mattered more than any single event.

Root credential use from external IP

Before: near miss, already present in telemetry.

After: context made root credential use decisive.

The agent separated weak signal from escalation-worthy evidence.

Wow factor: connect what rules see separately into one defensible finding.

© 2026 Cisco and/or its affiliates. All rights reserved.

Cisco Confidential

Technical deep dive: model, data, and service architecture

Source: CiscoLive slides 14, 24-28

One model family, layered by supervision

The same evidence layer scales from embeddings to task heads to agents.

Layer 0

Log Encoder

Universal embeddings for semantic search, clustering, and pruning.

Layer 1

Task heads

MITRE classification, key event summarization, anomaly ranking.

Layer 2

Agent service

A2A service surface for RCA, investigation, and product agents.

Layer 3

Graph reasoning

Entity relationships, paths, confidence, and blast radius.

Self-supervision gives breadth. Task-specific supervision gives precision. Agent and graph layers make it usable.

© 2026 Cisco and/or its affiliates. All rights reserved.

Cisco Confidential

Source: CiscoLive slides 24-25

How the log encoder works

It generates universal log embeddings that capture log semantics and syntax.

Input

Raw log lines

Structured fields

Incident or alert context

tokens

Transformer encoder

Embedding

Attention

FFN

Norm

vector

Output

Semantic vectors

Similarity scores

Cluster-ready representation

```
2026-03-03T12:00:01 checkout-service ERROR upstream timeout
2026-03-03T12:00:03 checkout-service WARN retry queue length=120
2026-03-03T12:00:07 checkout-service ERROR circuit breaker opened
```

Why encoder, not decoder

The encoder predicts masked tokens and learns reusable representations.

The downstream LLM reasons after evidence selection.

© 2026 Cisco and/or its affiliates. All rights reserved.

Cisco Confidential

Source: CiscoLive slides 30-31

Training corpus and pipeline

The moat is the log data Cisco and Splunk already live in.

300M

O11y logs / 1B+ raw lines

4M

security logs / 50M+ raw lines

14M

parameter model family target



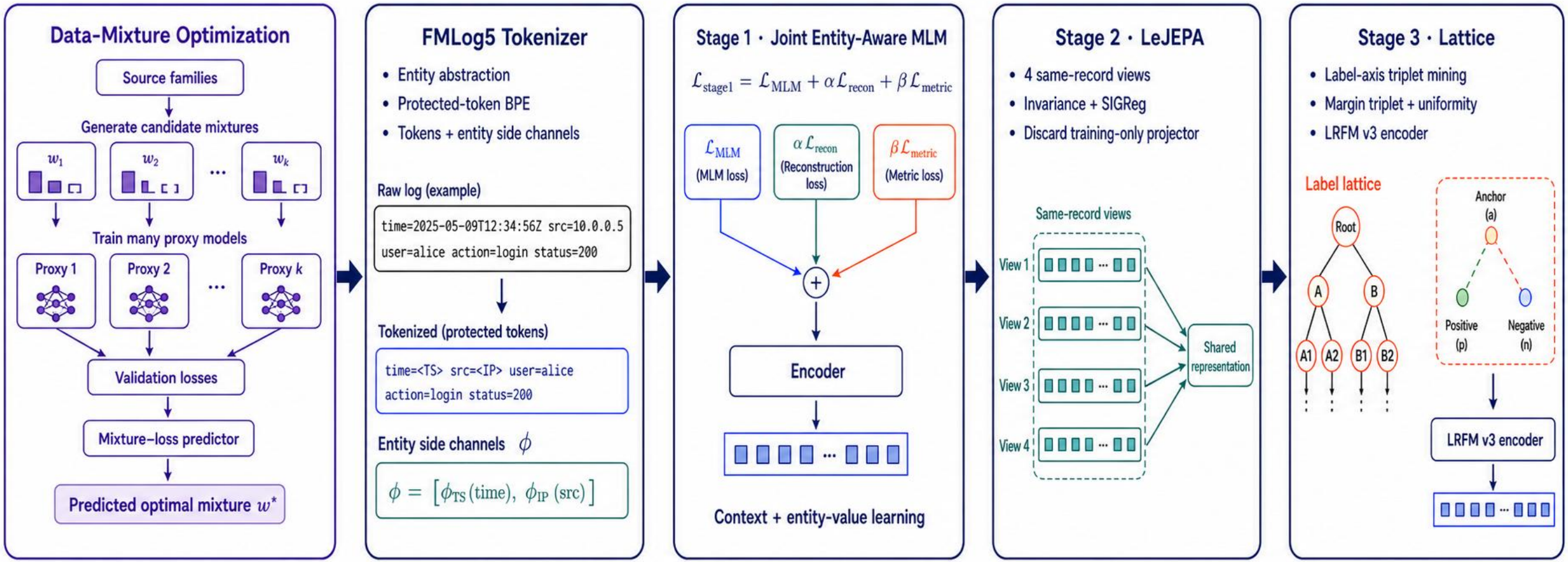
Key principle

Train on the logs we live in: Splunk-internal observability, security telemetry, and customer data by design.

© 2026 Cisco and/or its affiliates. All rights reserved.

Cisco Confidential

L0 Encoder Training Pipeline



© 2026 Cisco and/or its affiliates. All rights reserved.

Cisco Confidential

Source: CiscoLive slides 32-33

Dedupe and log-aware MLM

Remove redundancy while forcing the model to learn meaning, not surface artifacts.

Dedupe

SPL dedupe masks trivial variation such as timestamps.

Fuzzy dedupe removes semantically redundant event templates.

Net effect: reduce corpus size by 2-4 orders while preserving semantic diversity.

Log-aware masked language modeling

Selectively masks values, identifiers, and numeric quantities.

Preserves JSON keys, delimiters, whitespace, and syntax.

No labels required: pre-training is self-supervised before task labels appear.

Preserve diversity | **Mask semantics** | **Learn invariants**

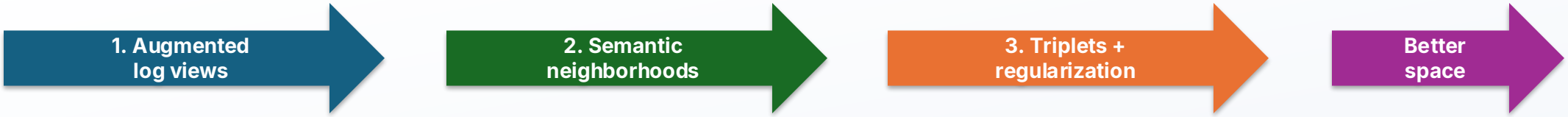
© 2026 Cisco and/or its affiliates. All rights reserved.

Cisco Confidential

Source: CiscoLive slides 34-35, 41-43

Self-supervision and contrastive learning

The full training recipe shapes embedding space for retrieval, clustering, and classification.



Pull together

- Different versions of the same log
- Logs close in the label lattice
- Events with shared meaning

Push apart

- Look-alike but semantically different logs
- Hard negatives from other entities
- Surface matches without causal meaning

Compound gain

- MLM -> self-supervision -> contrastive
- Win rate: 43% -> 56% -> 88%
- Strongest across retrieval, classification, clustering

© 2026 Cisco and/or its affiliates. All rights reserved.

Cisco Confidential

Source: CiscoLive slides 27, 44-46

Layer 1: task-specific learning

A lightweight task head turns the encoder into product-specific judgment.

MITRE classification

Multilabel classification across 132 techniques.

Domain-specific lattice triplet encoder predicted the exact set 84.6% of the time.

Key event summarization

Compress raw sequences into key events.

Give agents a semantic timeline instead of raw repetition.

Next-event prediction

Predict likely next log event from historical sequence.

Useful for proactive incident investigation and hypothesis generation.

2.86M
Sysmon logs in event-classification task

96%+
label noise filtered by relabeling

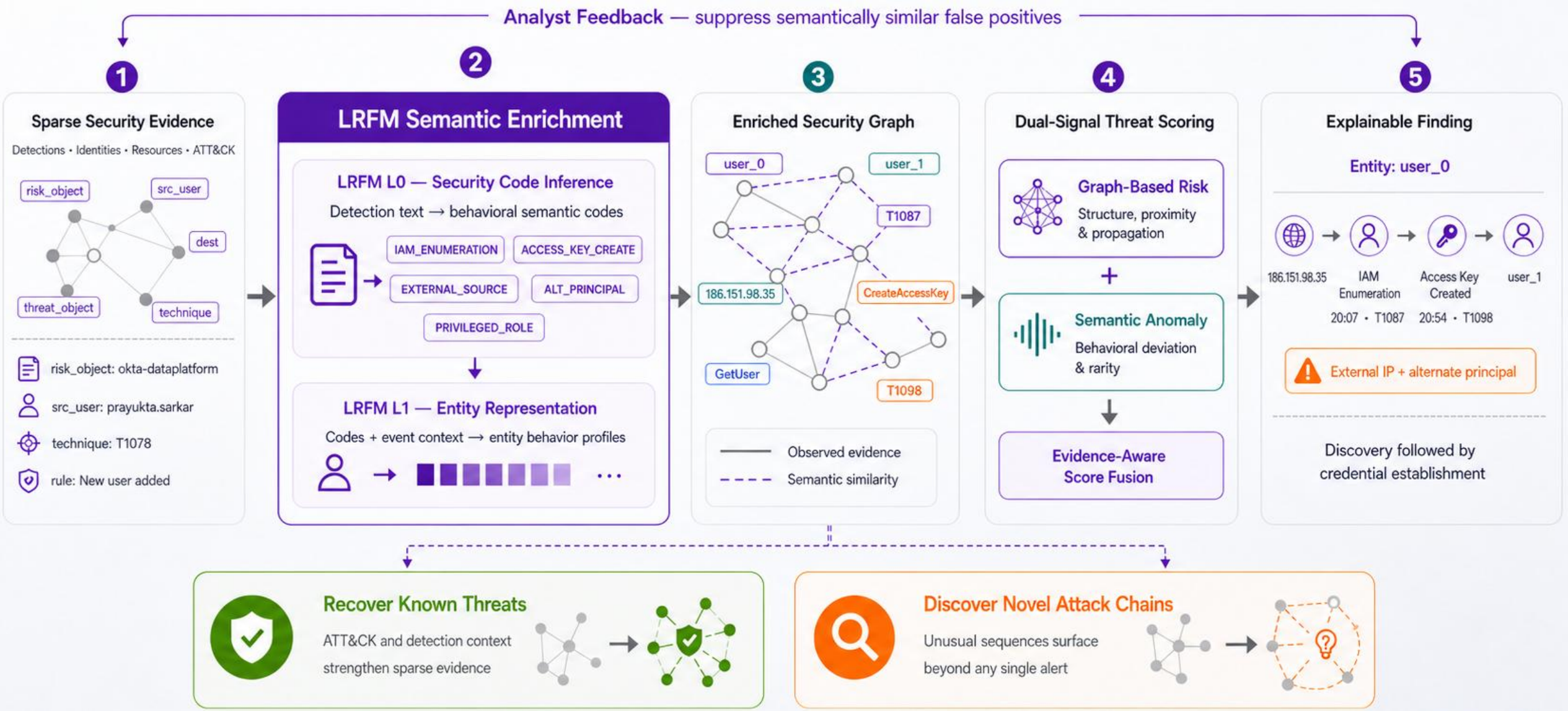
0.746
macro F1 at support >=50

40
fewest label failures

© 2026 Cisco and/or its affiliates. All rights reserved.

Cisco Confidential

AI Detection Engineer



© 2026 Cisco and/or its affiliates. All rights reserved.

Cisco Confidential

Source: CiscoLive slide 28 + AI deck

Layer 2: agentic log reasoning service

Expose the model as a deployable capability other agents can call.

Service surface

- Splunk Cloud deployable service
- Agent-to-agent communication
- Shared primitives for RCA and investigation

Product integrations

- O11y SRE Agent
- Enterprise Security AI Detection
- Search primitives and customer/partner agents

Safety posture

- Evidence links back to source logs
- Confidence and reason codes
- Graceful fallback when model path is unavailable

Foundation model value compounds when every agent can call the same log-native evidence layer.

© 2026 Cisco and/or its affiliates. All rights reserved.

Cisco Confidential

Source: CiscoLive slide 48 + pitch slide 35

What customers can expect

A log-native evidence layer powering AI agents across observability and security.

Broad coverage

Security, networking, and observability log types.

More sourcetypes improve model breadth over time.

Tenant-specific tuning

Safely fine-tune on customer data.

Learn topology, attributes, and incident habits that generic models miss.

Out-of-box agents

Use standalone or plug into existing agentic workflows.

Start with Log Pruner and AI Detection Agent use cases.

Put this in your domain: if the workflow finds the right logs before reasoning, LRFM can help.

© 2026 Cisco and/or its affiliates. All rights reserved.

Cisco Confidential

Thank you

Questions: Log Pruner, AI Detection Agent, and the LRFM technical path

Session ID TBD



splunk>
a CISCO company

Trust AI at scale