

SCOPE: Rethinking Attack Analysis as Uncertainty-Aware Security Reasoning under Incomplete Observations

JuYoung Lee

Gachon University

JeongMin Oh

Gachon University

SeonWoo Lee

Gachon University

TaeJin Lee

kinjecs0@gmail.com

Gachon University

Research Article

Keywords: Attack Chain Reconstruction, Endpoint Log Analysis, MITRE ATT&CK, Probabilistic Inference, Uncertainty Modeling, TTP Mapping

Posted Date: September 2nd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10595811/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

SCOPE: Rethinking Attack Analysis as Uncertainty-Aware Security Reasoning under Incomplete Observations

JuYoung Lee · JeongMin Oh · SeonWoo Lee · TaeJin Lee
(Corresponding author)

Received: date / Accepted: date

Abstract Understanding an attack’s overall picture is a key challenge in cybersecurity. In practice, log observations are incomplete and unreliable: activities go unrecorded, attackers remove or distort traces, and identical low-level actions signal different intents by context. Existing approaches face three fundamental limitations: (i) fixed analysis units fail to capture the varying scope and structure of attack behaviors, (ii) single-label decisions are brittle, so local errors break the analysis, and (iii) reliance on predefined rules or training distributions limits generalization to unseen attack compositions. We propose SCOPE, an uncertainty-aware framework for reconstructing attack behaviors under incomplete observations. SCOPE integrates three components: (1) *behavior-centric grouping* that forms semantically coherent units robust to missing or disordered events, (2) *distributional TTP mapping* that preserves ambiguity by maintaining candidate technique distributions instead of a single label, and (3) *Top-K Viterbi sequence inference* that

combines tactical and causal compatibility and stays robust to missing steps. Structural inference is deterministic and training-free—no attack sequences or transition frequencies are learned; the large language model (LLM) only generates auxiliary descriptions. On the OTRF Security-Datasets—chosen for its ground-truth attack flows and controlled log-loss stress tests—SCOPE achieves a Top-5 hit rate of 0.84 for tactics, techniques, and procedures (TTPs) and a technique longest-common-subsequence (LCS) of 0.68, outperforming event-, graph-, LLM-, and sequence-based baselines while staying robust under significant log loss. It also reconstructs coherent re-combinations of known MITRE ATT&CK techniques absent from the campaign library (coherent-novel rate 0.83). Uncertainty-aware modelling provides a practical foundation for robust attack-chain analysis.

Keywords Attack Chain Reconstruction · Endpoint Log Analysis · MITRE ATT&CK · Probabilistic Inference · Uncertainty Modeling · TTP Mapping

Corresponding author: TaeJin Lee

JuYoung Lee · SeonWoo Lee
Department of Information Security, Gachon University,
1342, Seongnam-daero, Sujeong-gu, Seongnam-si 13120,
Gyeonggi-do, Republic of Korea
E-mail: jylee0741@gmail.com (JuYoung Lee, ORCID 0009-
0001-8736-9251)
E-mail: dltjs5621@gachon.ac.kr (SeonWoo Lee, ORCID 0009-
0006-1094-922X)

JeongMin Oh · TaeJin Lee (corresponding author)
Department of Smart Security, Gachon University, 1342,
Seongnam-daero, Sujeong-gu, Seongnam-si 13120, Gyeonggi-
do, Republic of Korea
E-mail: ojmes5790@gmail.com (JeongMin Oh, ORCID 0009-
0008-7463-3966)
E-mail: kinjecs0@gmail.com (TaeJin Lee, ORCID 0000-0003-
3078-3459)

1 Introduction

Modern advanced persistent threats (APTs) unfold across multiple stages and leave behind large volumes of fine-grained endpoint logs spanning processes, files, registries, and network activity. Reconstructing the overall attack structure from such logs is central to effective threat detection and response, yet the logs themselves are inherently incomplete, noisy, and only partially ordered—both by system limitations and by adversarial manipulation [1, 2]—so analysts in production SOC’s see streams of local alerts but cannot recover the multi-stage intent that makes an incident actionable.

Under these conditions, we identify three fundamental challenges that remain unresolved by existing approaches.

(P1) Granularity gap between raw events and attack semantics. A single endpoint event—a process creation, a file write, a registry change—is too fine-grained to convey the intent of an attack step. The same event type can correspond to very different adversarial behaviors depending on context, and meaningful behaviors are realized as interleaved operations across multiple entities [3]. Current pipelines extract log sequences via fixed time windows or per-process boundaries, yet these units fail to align with the semantic boundaries of ATT&CK techniques, whose execution spans vary widely in duration and entity scope [4, 5]. At the other extreme, provenance graphs capture rich structural information but incur prohibitive construction cost and become structurally incomplete as soon as intermediate events are lost [1, 6]. What is missing is an *intermediate representation* that preserves semantically coherent units of activity while remaining robust to missing or reordered observations—no prior unit of analysis satisfies both simultaneously.

(P2) Brittle deterministic inference. Existing reconstruction methods commit to a single TTP assignment per analysis unit and chain them sequentially; a single misclassification propagates forward and breaks the entire attack narrative. Graph-based approaches suffer an analogous failure mode when intermediate causal edges are missing [6, 7]. LLM-based methods offer promising semantic reasoning, yet remain unreliable for structural inference due to hallucination and context-length constraints [8–10]. TTP sequence models—including HMM-based attack plan recognition [11] and deep-learning sequence predictors [12]—address the sequence dimension but assume the input is an already-classified TTP sequence, bypassing the semantic gap where most of the ambiguity lives. As a result, no existing method *jointly* resolves semantic ambiguity in mapping and structural uncertainty in sequencing. Decoupling the two lets local mapping errors propagate without recourse and leaves structural gaps beyond the reach of local evidence—yet incomplete observations demand precisely this joint recovery.

(P3) Dependence on the training distribution. Even if P1 and P2 were addressed in principle, existing systems remain bounded by the attack space they were trained or programmed to recognize. Current methods implicitly detect only the attack patterns covered by their predefined rules, training distributions, or known campaign libraries. Rule-based detectors cannot recognize behaviors outside their rule scope [13];

learning-based approaches generalize only within their training distribution [14]; and RAG-based mapping systems such as RAM are bounded by the rule set available at deployment [15]. A framework unable to generalize to *unseen compositions of known techniques* fails precisely where adversarial novelty matters most—at the intersection of known primitives arranged in unknown orderings.

We address these three challenges with SCOPE (Security Reasoning under Partial Observability), an *uncertainty-aware* framework that reframes endpoint log analysis as *probabilistic attack chain reconstruction*. SCOPE proceeds in three stages. First, raw logs are transformed into semantically coherent *behavior groups* by aggregating events around anchor actions using temporal-locality, process-lineage, and shared-entity criteria. Unlike prior behavior abstraction aimed at semantic summarization [3], our grouping is explicitly designed as the input unit for downstream probabilistic inference. Second, each group is mapped to a *distribution* over candidate ATT&CK techniques via embedding similarity—preserving local uncertainty rather than collapsing it into a single label—and thereby allowing the mapping to deviate from the anchor’s rule prior when the semantics warrant it. Third, the groups are assembled into a global chain by *probabilistic sequence inference*: consecutive behaviors are scored through a *tactical-causal transition function* combining kill-chain flow and causal lineage—derived from domain-knowledge priors rather than dataset-specific frequencies, so the model generalizes to unseen compositions—and a Top- K Viterbi recurrence recovers the most plausible sequence. When intermediate events are missing, the surviving groups are linked in observed order and the tactical prior discounts the resulting cross-tactic jumps rather than breaking the chain. A central design principle is the *separation of semantic abstraction and structural inference*: the LLM participates only in description generation, while every decision is made by a deterministic, training-free probabilistic model. This confines hallucination risk to an auxiliary step whose output is subsequently validated by downstream scoring. The reconstructed chain is the artifact downstream triage and scoping workflows consume: a scored, ordered hypothesis about the underlying incident in place of a stream of isolated alerts.

Contributions.

- **Uncertainty-aware behavior-centric reconstruction (C1).** We propose a unified framework coupling behavior-group abstraction with a Top- K Viterbi recurrence that jointly optimizes local TTP likelihoods and tactical-causal transition compat-

ibility, preserving global chain continuity across observed-evidence gaps with the LLM strictly confined to description generation.

- **Training-free generalization to unseen attack compositions (C2).** We show that distributional TTP mapping and compatibility-based (rather than learned) transition scoring—neither of which requires labeled attack sequences—support recovery of recombinations and tactical orderings of known ATT&CK techniques absent from any training corpus, quantified through rule-prior deviation and chain-level novelty metrics.
- **Multi-tier controlled evaluation (C3).** We evaluate against four baseline families (event-level, graph-based, LLM-based, TTP-sequence) on a curated benchmark of 35 OTRF scenarios; ablation studies show that the layered design is necessary and no single component is sufficient.

Evaluation strategy. We use OTRF not as a substitute for enterprise deployment but as a setting where ground-truth attack flows and controlled log-loss stress tests are both available, making reconstruction under partial observability directly measurable; production-scale, multi-host, and benign-only validation are complementary steps discussed in Section 8.

2 Background

2.1 ATT&CK Framework and Endpoint Telemetry

The MITRE ATT&CK framework [16] catalogues adversary behavior as *Tactics*, *Techniques*, and *Procedures* (TTPs). *Tactics* are high-level adversarial objectives (e.g., PERSISTENCE, CREDENTIAL ACCESS, EXFILTRATION); *Techniques* are the concrete actions that realize a tactic, each identified by a technique ID (TID); and many techniques are further refined into *sub-techniques*. For instance, under the tactic CREDENTIAL ACCESS, the technique *OS Credential Dumping* (T1003) has sub-techniques such as *LSASS Memory* (T1003.001), where the parent (T1003) and the child (T1003.001) describe the same behavior at coarser and finer granularity — an ambiguity that directly affects how reconstructions should be scored (Section 7.1). A *procedure* is a specific in-the-wild realization of a technique. Modern security operations rely on this taxonomy for behavior-centric detection, but mapping raw telemetry to ATT&CK techniques remains challenging: a single technique may be implemented through diverse operations, and a single operation may match multiple techniques depending on context.

Endpoint systems produce fine-grained telemetry through Sysmon and Windows Event Logs, covering process creation, file access, registry changes, and network connections. Despite this granularity, individual events present a severe *semantic gap*—they are too fine-grained to convey adversarial intent, and malicious actions are interleaved with thousands of benign operations [3, 17].

Limitations of current analysis pipelines. Figure 1 illustrates the typical pipeline adopted by recent endpoint security systems: raw logs are normalized, segmented into sequences, matched against a TTP knowledge base via embedding similarity, and interpreted by an LLM for attack chain reasoning [4, 8, 10]. Although this pipeline has matured rapidly, its sequence extraction stage introduces three structural limitations that propagate through all downstream components:

- ① Fixed time-window or pivot-based approaches cannot accommodate the varying duration and entity scope of different ATT&CK techniques, so semantic context is lost or diluted by noise [4].
- ② Within a single process, heterogeneous activity types (file, network, registry, DLL) lack a principled filtering strategy to distinguish attack-relevant operations from routine ones [13].
- ③ Coordinated attack behaviors spanning multiple processes—the hallmark of real-world intrusions—have no established grouping mechanism [4, 9].

These limitations motivate the behavior-centric grouping introduced in Section 5.2.

2.2 Observation Model

We treat endpoint logs as imperfect observations of the underlying attack process, characterized by four properties. *Incomplete observations*: under heavy load, audit collectors may drop a large fraction of events, and buffered records can be erased before durable storage [1]. *Ambiguous event semantics*: the same event type maps to different behaviors depending on context [3]. *Fragmented evidence*: EDR detections do not explicitly link related alerts, and short retention windows further limit context [17]. *Unreliable temporal ordering*: timestamps can be changed, and the observed order may differ from the true order of execution [2].

Taken together, these properties imply that any reconstruction method must operate under uncertainty in both event interpretation and event connectivity.

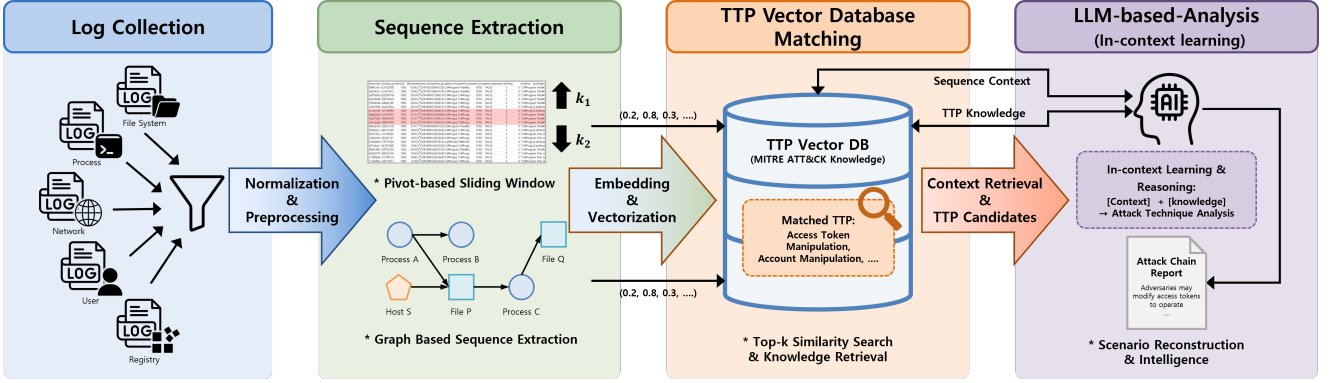


Fig. 1 Typical SOTA pipeline for endpoint attack reconstruction: logs are normalized, segmented (sliding-window or graph-based), matched against a TTP vector DB via top- k search, and interpreted by an LLM. Three structural limitations at the segmentation stage (①–③) propagate downstream.

Table 1 Capability-level comparison of SCOPE with seven representative research families. $\checkmark$ = fully addressed; Δ = partially; $\times$ = not addressed; N/A = not applicable. † HMM/Viterbi assumes an already-classified TTP sequence, bypassing semantic abstraction. * *Missing-log robustness* holds under the surviving-anchor assumption (Section 8). *Unseen-composition support* = recombinations of known techniques in unobserved orderings, not discovery of new behavior classes.

Family (representative approach)	Semantic abstraction	Uncertainty modeling	Missing-log robustness	Chain reconstruction	Unseen-composition support	LLM-dependency control
Event / Rule-based (Sigma)	$\times$	$\times$	$\times$	$\times$	$\times$	N/A
Behavior abstraction (WATSON)	$\checkmark$	$\times$	Δ	$\times$	$\times$	N/A
Graph-based (Provenance)	$\checkmark$	$\times$	$\times$	$\checkmark$	$\times$	N/A
HMM / Viterbi (sequence inference)	$\times^\dagger$	$\checkmark$	$\times$	$\checkmark$	$\times$	N/A
Deep-learning (GNN / LSTM)	Δ	$\checkmark$	$\times$	$\checkmark$	Δ	N/A
LLM-based reasoning (SHIELD)	$\checkmark$	Δ	$\times$	Δ	Δ	$\times$
CTI / Knowledge graph (AttackKG)	$\checkmark$	$\times$	$\times$	Δ	$\checkmark$	N/A
SCOPE (ours)	$\checkmark$	$\checkmark$	$\checkmark^*$	$\checkmark$	$\checkmark$	$\checkmark$

3 Related Work

We organize existing approaches into four representative families and position SCOPE relative to each across the six capability axes of Table 1.

3.1 Event- and Behavior-level Detection

Rule-based systems such as Sigma [18] match predefined signatures against log fields, and extensions like LogDLR [5] learn domain-invariant representations for cross-system anomaly detection. WATSON [3] aggregates audit-log events into behavior-level units through contextual semantics, addressing the semantic poverty of raw events. Both lines, however, work in narrow scope: event-level methods cannot recover progression when intermediate logs are missing, and WATSON stops at grouping—it does not model TTP uncertainty or reconstruct inter-behavior structure. SCOPE keeps behavior-level abstraction but treats groups as input units for downstream probabilistic chain inference over the full ATT&CK technique set.

3.2 Graph and LLM-based Reasoning

Provenance graph methods model audit logs as directed graphs of system entities and causal dependencies, with applications spanning threat hunting [6, 19], graph segmentation [4], context-aware reduction [7], and subgraph anomaly detection with LLM explanation [20]. Graphs provide the richest structural representation but require sufficient edge completeness — once intermediate events are lost, path-based traversal fails [1] — and full-graph construction incurs significant overhead [4, 6]. A complementary line layers LLMs on top of structural preprocessing for semantic reasoning: SHIELD [8], OMNISEC [9], MalGTA [10], APT-LLM [14], and RAM [15]. These approaches improve semantic abstraction but inherit upstream structural limitations: when segmentation is incomplete the LLM receives a distorted view and can produce structurally incorrect narratives [8, 21]. SCOPE retains structural grounding by incorporating causal lineage as a transition feature—a missing edge degrades the score rather than disconnecting the chain—and confines the LLM solely to description generation.

3.3 Sequence Inference: HMMs and Deep Learning

HMMs have been used for multi-step attack plan recognition from IDS alerts [11], and extended with ATT&CK-derived parameters for probabilistic attack sequence generation [22]. Deep-learning extensions include DeepOP [12], which combines ontology reasoning with technique-level sequence prediction, and DeepAG [23], which couples a Transformer encoder with a bi-directional LSTM over log-vector sequences derived from dataset-specific log templates. These models offer principled sequential reasoning, but they share two structural limitations. First, HMMs and DeepOP assume the input is an already-classified TTP sequence, bypassing the semantic gap between raw telemetry and TTP labels — precisely where noise, ambiguity, and data loss are most severe. DeepAG operates one level lower (over log-token sequences) but is tied to the log-template vocabulary on which it was trained. Second, both lines learn transition statistics from training corpora and remain distribution-dependent, limiting generalization to unseen TTP compositions. SCOPE preserves the probabilistic-inference spirit of HMMs while closing the semantic gap through behavior-group abstraction, and replaces frequency-based transitions with compatibility-based scoring from domain-knowledge priors.

3.4 CTI and Knowledge-Graph Reasoning

A complementary line builds structured knowledge graphs from cyber threat intelligence (CTI) reports to reason about attack techniques and campaigns. AttackKG, for example, constructs technique-level knowledge graphs by aligning attack behaviors extracted from unstructured threat reports [24]. Because knowledge-graph reasoning composes naturally over known primitives, such systems are particularly strong at capturing novel attack compositions, but they ground their analysis in textual reports rather than endpoint telemetry — they do not reconstruct chains directly from observed logs, do not model observation uncertainty, and do not handle missing events. SCOPE is complementary: it operates end-to-end on raw endpoint logs under explicit partial-observability assumptions while inheriting the compositional-novelty benefits via compatibility-based transition scoring.

3.5 Summary and Positioning

Collectively, the four families above show that semantic abstraction, uncertainty modeling, structural chain reconstruction, missing-log robustness, unseen-composition support, and LLM-dependency control have each been addressed individually, but never jointly. SCOPE occupies this gap by unifying behavior-group abstraction, distributional TTP mapping, tactical-causal transition compatibility, and a Top- K Viterbi recurrence, under a strict separation between LLM-based description generation and deterministic structural inference (Table 1). The contribution is the joint formulation: a regime in which no individual mechanism alone is sufficient (Sections 7.3 and 7.5).

We acknowledge that pairing an LLM with a deterministic scaffold is, in itself, an increasingly common pattern, including in log analysis [25]. SCOPE’s novelty is not the pairing per se but *what the deterministic layer models and how the LLM is confined*, along three axes that prior provenance-, sequence-, and LLM-scaffold systems do not provide jointly: (i) *uncertainty propagation* — each behavior carries a *distribution* over candidate techniques rather than a single label, so a locally ambiguous unit is resolved by global context instead of an early hard commitment; (ii) *missing observations as discount, not disconnection* — a dropped event leaves fewer groups whose surviving transition is merely down-weighted (the tactical prior discounts the cross-tactic jump, the set-overlap causal term softens) rather than broken, in contrast to provenance traversal that fails once an intermediate edge is lost; and (iii) *learning-free compatibility* — transitions are scored from kill-chain priors rather than corpus-learned frequencies, which is what lets SCOPE generalize to attack compositions absent from any training set. The LLM contributes only text descriptions and never participates in scoring or search (the swap ablation of Section 7.5 confirms robustness to model choice).

4 Problem Formulation

Building on the observation model (Section 2.2) and the limitations identified in Section 3, we formalize attack-chain reconstruction under incomplete observations.

Let $\mathcal{L}_{obs} = \{e_1, \dots, e_n\}$ denote the observed log events, where each e_i captures a system-level action. $\mathcal{L}_{obs}$ is an incomplete and noisy projection of an underlying sequence $\mathcal{L}_{true}$. Given $\mathcal{L}_{obs}$, the goal is to recover an attack chain

$$(\mathcal{B}, S) = (\{b_1, \dots, b_T\}, (t_1, \dots, t_T)),$$

where $\mathcal{B}$ is a sequence of semantically coherent behavior groups b_j and S is the TTP sequence assigning to each b_j a MITRE ATT&CK technique t_j , satisfying:

- **semantic consistency**: each behavior aligns with a plausible activity pattern;
- **structural plausibility**: a globally consistent ordering is recovered without requiring fully observed causal edges; and
- **robustness to missing observations**: non-adjacent surviving behaviors can be connected even when intermediate evidence is absent, filtered out, or too weak to form a reliable group.

We use *missing-step handling* throughout to denote preserving continuity across gaps in the observed behavior sequence: SCOPE does not fabricate unobserved TTP labels but connects non-adjacent surviving behavior groups through a penalized transition when the intervening evidence is absent, weak, or filtered out.

On terminology. The reconstructed object $(\mathcal{B}, S)$ is an *ordered sequence of behavior groups*—each binding the concrete events and entities (processes, files, registry keys, network endpoints) of a step together with its technique label—not a bare list of technique IDs. We call it an *attack chain*: it conveys what happened, on which entities, and in what order, with each group retaining its events for an analyst to drill into. Emitting a complete entity-level causal graph (fixed entry point, every intermediate node, fully connected paths) is deliberately out of scope, because requiring complete causal connectivity is precisely the assumption that fails under missing logs (Section 3); SCOPE instead recovers the technique-level progression that stays well-defined under partial observability while preserving per-step entity context.

This reframes attack analysis as *behavior-level abstraction* coupled with *probabilistic reconstruction under partial observability*, eliminating the need for complete event coverage, explicit graph connectivity, or strictly deterministic reasoning. Any solution that decouples these three requirements satisfies at most two of them under incomplete observations—motivating a unified framework in which they are addressed jointly.

5 Detailed Design

Following the observation model of Section 2.2, SCOPE is organized around two complementary pillars: (i) *behavior-centric semantic abstraction*, an intermediate representation between raw events and ATT&CK techniques that preserves meaning under missing or reordered events; and (ii) *uncertainty-aware*

probabilistic inference, which maintains a distribution over candidate techniques per step and recovers a globally consistent sequence by jointly exploiting local evidence and tactical-causal transition compatibility. Two operational properties emerge as direct consequences of these pillars: *rule-prior deviation* (Section 5.3), where semantic mapping can assign a behavior group to a technique that differs from its anchor rule’s tagged prior, and *chain-level novelty* (Section 5.5), where compatibility-scored sequences can correspond to compositions absent from any known-campaign library.

5.1 System Overview and Running Example

Pipeline. Figure 2 shows SCOPE’s pipeline, mapping directly to the requirements of Section 4: (1) *behavior grouping*, aggregating events around anchor actions via temporal-locality, process-lineage, and shared-entity criteria (Section 5.2); (2) *representation and distributional mapping*, deriving a distribution $P_{\text{ttp}}(\cdot \mid b_j)$ over ATT&CK techniques from a natural-language description of each group (Section 5.3); and (3) *probabilistic sequence inference*, which scores consecutive behaviors along tactical and causal axes (Section 5.4) and recovers the most plausible global sequence with a Top- K Viterbi recurrence (Section 5.5). Stages 1–2 secure *semantic consistency*, while Stage 3 enforces *structural plausibility* and *robustness to missing observations*.

Design separation. SCOPE separates *semantic abstraction* (LLM-driven, stochastic) from *structural inference* (deterministic scoring): the LLM only produces descriptions and never decides the chain, confining hallucination to an auxiliary step that similarity retrieval and transition scoring validate.

Running example. A fragment of *msf_record.mic* (one intermediate DISCOVERY step assumed unobserved) threads through this section; full walkthrough in A.

Notation. We write $\mathcal{L}_{\text{obs}} = \{e_1, \dots, e_n\}$ for the observed log, a_j and b_j for an anchor event and its surrounding behavior group, d_j for the natural-language description of b_j , $\mathcal{T}$ for the MITRE ATT&CK technique set, and $P_{\text{ttp}}(t \mid b_j)$ for the per-group candidate distribution over $\mathcal{T}$. The fused transition score $P_{\text{trans}}(t_i \rightarrow t_j)$ is composed from tactical and causal terms $P_{\text{tac}}, P_{\text{cau}}$. The Top- K Viterbi recurrence carries a forward score $\delta[\cdot][\cdot]$ and back-pointer $\psi[\cdot][\cdot]$; its output is the inferred TTP sequence S^* with novelty score $n(S^*)$.

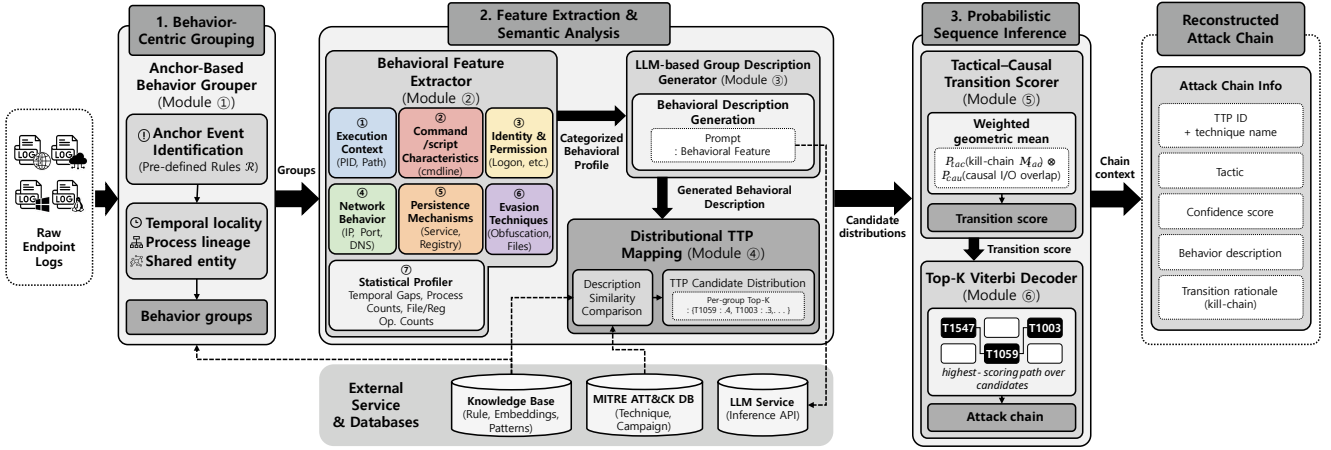


Fig. 2 Overview of SCOPE. Raw endpoint logs pass through three stages: (1) *behavior-centric grouping* around rule anchors; (2) *representation and distributinal TTP mapping*, where structured features drive a *confined* LLM description and a per-group distribution over candidate techniques; and (3) *probabilistic sequence inference*, where a tactical-causal transition scorer feeds a Top-K Viterbi decoder. The result is a reconstructed attack chain whose steps carry technique, tactic, confidence, a behavior description, and a transition rationale; chain-level context (novelty against the known-campaign library $\mathcal{C}$) is provided for contextualization only, not as attribution. The LLM is confined to description generation and never participates in scoring or search; external resources are read-only.

5.2 Behavior-Centric Log Grouping

Motivation. A single audit event is rarely sufficient to convey adversarial intent: the same process creation can correspond to benign administration or to a discovery step, and meaningful behavior is typically realized as a cluster of interleaved operations across processes, files, registry keys, and network endpoints [3]. Directly labeling individual events therefore fragments semantics and is also brittle against missing intermediates. We instead construct an intermediate unit—a *behavior group*—that is both semantically coherent and robust to holes, by aggregating related events around an identified anchor action.

Definitions. An *anchor event* $a \in \mathcal{L}_{obs}$ is an event whose type and attributes match a template indicative of a potentially attack-relevant action (e.g., a credential dump via `comsvcs.dll`, a PowerShell download cradle, or a registry `Run` key modification). Anchors are identified by a lightweight rule set $\mathcal{R}$ that performs *triage*, *not classification*: it decides only *where* to look and delegates the technique decision to the semantic mapping of Section 5.3. This separation is deliberate — unlike rule-based SIEM/EDR, where a rule both flags *and* labels an event so that a mis-specified rule ends the analysis, the anchor merely surfaces a candidate region and every technique is re-decided downstream against the full ATT&CK catalogue (36% of groups map to a technique other than their anchor’s prior; Section 7.4). Casting a broad, recall-first net over suspicious activity is the standard first step of triage-based analysis, and confining rules to that role is what lets SCOPE rein-

terpret rule-flagged activity rather than inherit a rule’s decision. *Supporting events* are events that enrich the context of an anchor by supplying process lineage, command arguments, or shared-artifact interactions. A *behavior group* $b_j \subseteq \mathcal{L}_{obs}$ is the union of an anchor and the supporting events admitted by the criteria below.

Grouping criteria. For each anchor a , three criteria jointly determine its supporting set:

- *Temporal locality*: events within a window τ_t around a . Used as a recall filter rather than as the causal backbone, and therefore tolerant of non-monotonic timestamps [2].
- *Process lineage (h -hop)*: events whose process identifier lies within h process-tree hops of a . Lineage links are preferred over timestamps precisely because they are less manipulable in practice.
- *Shared entities*: events that touch at least one entity (file path, registry key, network endpoint, or named pipe) also touched by a , including cross-field matches such as *anchor-dropped file* $\leftrightarrow$ *later-executed image*. This preserves cross-artifact continuity beyond process scope and is the key ingredient that allows SCOPE to assemble groups when lineage evidence is broken by registry-mediated persistence, drop-then-execute patterns, or log-clearing gaps.

These three criteria correspond to orthogonal signals of behavioral coherence—temporal co-occurrence, control-flow lineage, and data-flow artifact sharing—so that losing any single signal (e.g., a missing intermediate event or a manipulated timestamp) does not elimi-

Algorithm 1 Behavior-Centric Log Grouping

Require: log $\mathcal{L}_{obs}$, anchor rules $\mathcal{R}$, window τ_t , depth h
Ensure: behavior group set $\mathcal{B}$

```

1:  $\mathcal{B} \leftarrow \emptyset$ ;  $\mathcal{A} \leftarrow \text{MATCHANCHORS}(\mathcal{L}_{obs}, \mathcal{R})$ 
2: for all  $a \in \mathcal{A}$  do
3:    $W \leftarrow \{e \in \mathcal{L}_{obs} : |t(e) - t(a)| \leq \tau_t\}$ 
4:    $L_a \leftarrow \text{LINEAGE}(a, W, h)$ 
5:    $E_a \leftarrow \{e \in W : \text{ENTITIES}(e) \cap \text{ENTITIES}(a) \neq \emptyset\}$ 
6:    $b \leftarrow \{a\} \cup L_a \cup E_a$ 
7:    $\mathcal{B} \leftarrow \mathcal{B} \cup \{b\}$ 
8: end for
9:  $\text{MERGEOVERLAPPING}(\mathcal{B}) \triangleright$  union-find on shared anchors
10: return  $\mathcal{B}$ 
```

nate the grouping evidence entirely. The full procedure is given in Algorithm 1.

Robustness to temporal and structural noise. The three criteria are intentionally *disjunctive*: an event joins a group whenever process lineage *or* entity sharing holds within the temporal window, so a dropped intermediate weakens at most one criterion while the other preserves coherence — e.g., the shared-entity criterion bridges registry-mediated persistence and drop-then-execute sequences where pure lineage grouping would lose the continuation. This addresses **P1**: behavior groups give a technique-aligned unit that stays coherent under dropped or reordered events.

5.3 Behavior Representation and TTP Candidate Distribution

From groups to behavioral descriptions. Given a behavior group b_j , SCOPE extracts a structured feature vector $x_j = \text{FeatureExtract}(b_j)$ summarizing the group along seven orthogonal axes (execution context, command & script, identity & privilege, network, persistence, evasion, and statistical profile; full schema in C). A lightweight generator then produces a concise natural-language description $d_j = \text{Describe}(x_j)$ used purely as a semantic bridge: it does not directly determine a TTP label (Section 5.1).

Candidate TTP distribution. Rather than committing to a single label, SCOPE constructs a distribution over candidate techniques by scoring the similarity of d_j against the reference description d_t of every technique $t \in \mathcal{T}$:

$$P_{\text{ttp}}(t \mid b_j) = \frac{\exp(\beta \cdot \text{sim}(d_j, d_t))}{\sum_{t' \in \mathcal{T}} \exp(\beta \cdot \text{sim}(d_j, d_{t'}))}, \quad (1)$$

where $\text{sim}(\cdot, \cdot)$ is cosine similarity in an embedding space (we use `all-mpnet-base-v2` [26, 27]) and $\beta > 0$ is a temperature parameter. We retain only the Top- K candidates per group, absorbing the residual mass into a single symbol. These are normalized plausibility

scores, not calibrated posteriors (calibration is left for follow-up work, Section 8).

Confidence margin. The *confidence margin* $m_j = P_{\text{ttp}}(t^{(1)} \mid b_j) - P_{\text{ttp}}(t^{(2)} \mid b_j)$ is a local uncertainty signal consumed by the inference module: small margins flag semantically ambiguous groups to be resolved by sequence-level evidence rather than a forced local decision — directly addressing **P2**.

Consequently SCOPE is not bound to a fired anchor’s technique prior: the mapping step may return any ATT&CK technique whose reference best matches d_j , so a group whose top-1 differs from its anchor template (margin $m_j \geq \mu$) is a *rule-prior deviation* — the semantic layer re-interpreting rule-flagged activity, which is what supports the *unseen composition of known techniques* case. As in any triage-based pipeline, coverage is ultimately bounded by anchor recall, but this failure stays localized — an anchor-invisible behavior yields a missed group, not a corrupted chain (Section 8).

5.4 Tactical-Causal Transition Compatibility

Rationale. Selecting the highest-probability candidate independently for each group is brittle: an ambiguous local decision may be decisively resolved by context that appears one or two steps later. SCOPE therefore scores every ordered pair of candidate techniques (t_i, t_j) along two complementary axes—tactical and causal. These scores reflect *compatibility* with a plausible attack progression rather than empirical frequencies learned from a fixed attack corpus, so the transition model supports chain-level novelty without transition-frequency retraining; the tactical weights (Table 10) encode kill-chain semantics rather than dataset-specific tuning, and SCOPE’s accuracy is robust to $\pm 30\%$ perturbation of every matrix weight (D).

Tactical flow compatibility. Let $\tau(t)$ denote the ATT&CK tactic associated with technique t , and let $\mathcal{K}$ be the set of all ATT&CK tactics. We define

$$P_{\text{tac}}(t_i \rightarrow t_j) = M_{\text{tac}}[\tau(t_i), \tau(t_j)], \quad (2)$$

where $M_{\text{tac}} \in \mathbb{R}_{\geq 0}^{|\mathcal{K}| \times |\mathcal{K}|}$ is a compatibility matrix whose entries are pre-defined by transition type rather than learned from corpus frequencies: primary adjacent progressions (e.g., Execution $\rightarrow$ Persistence) receive the highest weight, post-exploit cycles among CREDENTIAL ACCESS, DISCOVERY, and LATERAL MOVEMENT are rewarded slightly less, long forward skips are discounted, and transitions that reverse across the kill chain or undo IMPACT/EXFILTRATION are effectively

forbidden. These transition types follow the forward progression codified in the intrusion kill chain [28] and the MITRE ATT&CK tactic structure [16], and are consistent with the transition regularities documented empirically across real multi-stage incidents [29]. The full transition-type catalogue is tabulated in B (Table 10), and the complete materialized 14×14 matrix M_{tac} is given explicitly in Table 2.

Causal lineage. Each technique t is associated with input/output entity-type signatures $\text{In}(t), \text{Out}(t)$ drawn from {process, file, registry, network, service, user} — the artifact types a step consumes and produces. We then score

$$P_{\text{cau}}(t_i \rightarrow t_j) = \frac{|\text{Out}(t_i) \cap \text{In}(t_j) \cap \mathcal{E}(b_i, b_j)|}{|\text{In}(t_j)| + \varepsilon}, \quad (3)$$

where $\mathcal{E}(b_i, b_j)$ is the set of entity *types* observed to be shared between the two groups and ε is a small constant. Unlike strict graph traversal, missing an individual edge only reduces $\mathcal{E}$ rather than disconnecting the chain.

Adaptive fusion. We combine the two components as a weighted geometric mean,

$$P_{\text{trans}}(t_i \rightarrow t_j) = \exp\left(\sum_{k \in \{\text{tac}, \text{cau}\}} w_k(b_i, b_j) \cdot \log P_k(t_i \rightarrow t_j)\right), \quad (4)$$

which penalizes any axis scoring near zero. The weights w_k are *data-adaptive*: w_{cau} is modulated by the entity-coverage ratio $\rho = |\mathcal{E}(b_i, b_j)|/|\text{In}(t_j)|$ (subject to $\sum_k w_k = 1$), and is down-weighted—redistributing mass to w_{tac} —when logging is lossy.

5.5 Probabilistic Sequence Inference

Objective. Given an ordered sequence of behavior groups $\mathcal{B} = \{b_1, \dots, b_T\}$ together with the per-step distributions of Section 5.3 and the pairwise compatibilities of Section 5.4, we seek the TTP sequence $S^* = (t_1^*, \dots, t_T^*)$ that maximizes joint plausibility:

$$S^* = \arg \max_S \prod_{s=1}^T P_{\text{ttp}}(t_s | b_s) \cdot \prod_{s=2}^T P_{\text{trans}}(t_{s-1} \rightarrow t_s)^\alpha, \quad (5)$$

where $\alpha \in (0, 1]$ is the transition-to-emission weight (default $\alpha = 0.5$, calibrated via the held-out sweep of D). Because the number of candidate sequences grows exponentially, we solve the objective via a Top- K Viterbi recurrence that retains only the K most plausible candidates at each step.

Recurrence. Let $\delta[s][t]$ be the best joint score of any sequence ending with technique t at position s . We define

$$\delta[s][t] = P_{\text{ttp}}(t | b_s) \cdot \max_{t' \in \text{Top}_K(b_{s-1})} \left[\delta[s-1][t'] \cdot P_{\text{trans}}(t' \rightarrow t)^\alpha \right], \quad (6)$$

which links each behavior group to its immediate predecessor in observed time order; back-pointers $\psi[s][t]$ record the predecessor t' for trajectory recovery. In log-space, $\log \delta[s][t] = \log P_{\text{ttp}}(t | b_s) + \max_{t'} [\alpha \log P_{\text{trans}}(t' \rightarrow t) + \log \delta[s-1][t']]$. Crucially, missing observations require no dedicated recovery mechanism. Dropped events simply yield *fewer* behavior groups, and the recurrence links the surviving groups in their observed order; a gap manifests as a single transition whose endpoints jump across the kill chain (e.g., EXECUTION directly to COLLECTION). The tactical matrix M_{tac} already *discounts* such forward jumps through its off-diagonal forward entries (Table 2) rather than breaking the chain. Continuity under partial observability is therefore preserved by down-weighting incongruous transitions between surviving groups, not by inserting latent ones—directly addressing the *robustness to missing observations* requirement of Section 4.

The recurrence is computed in standard Top- K Viterbi fashion: at each behavior position, only the K highest-scoring partial paths into each Top- K candidate are retained. The total cost is $O(T \cdot K^2)$ time and $O(T \cdot K)$ memory, where K is the beam width. The full procedure is given in Algorithm 2, with a running-example lattice (Top-3 columns) in Figure 3.

Emission-confidence bypass. When the emission is already highly confident— $\text{sim}(d_j, d_{t^{(1)}}) \geq \tau_{\text{emit}}$ for the top-ranked technique $t^{(1)}$ —structural rerank by a noisier P_{trans} can substitute an adjacent candidate that fits the chain better *locally* but is less accurate *per group*. We therefore adopt an asymmetric-trust selection rule: after the DP returns its globally optimal path, each position s is replaced by the emission argmax whenever $\text{sim}(d_s, d_{t_s^{(1)}}) \geq \tau_{\text{emit}}$, and otherwise retains the Viterbi pick. The threshold $\tau_{\text{emit}} = 0.75$ (p_{75} of top-1 similarities on OTRF) restricts the bypass to the $\approx 25\%$ most confident groups. This is a post-processing step rather than a modification of the DP objective: when the bypass fires, the final chain is not necessarily the exact maximizer of Equation 6, and we report its ablation separately in Section 7.5. For chain-score accounting, bypassed positions are treated as neutral, so novelty measurements remain driven by the uncertain positions where structural reasoning actually fires.

Table 2 Complete explicit tactical transition matrix M_{tac} (row = source tactic, column = destination tactic), materialized from the rule catalogue in Table 10. Tactics: RE Reconnaissance, RD Resource Development, IA Initial Access, EX Execution, PE Persistence, PV Privilege Escalation, DE Defense Evasion, CA Credential Access, DI Discovery, LM Lateral Movement, CO Collection, C2 Command and Control, EF Exfiltration, IM Impact.

	RE	RD	IA	EX	PE	PV	DE	CA	DI	LM	CO	C2	EF	IM
RE	0.50	1.00	0.50	0.30	0.30	0.80	0.70	0.30	0.30	0.30	0.30	0.30	0.30	0.30
RD	0.10	0.50	1.00	0.50	0.30	0.80	0.70	0.30	0.30	0.30	0.30	0.30	0.30	0.30
IA	0.10	0.10	0.50	1.00	0.50	0.80	0.70	0.30	0.30	0.30	0.30	0.30	0.30	0.30
EX	0.05	0.10	0.10	0.50	1.00	1.00	1.00	1.00	0.30	0.30	0.30	0.30	0.30	0.30
PE	0.05	0.05	0.10	0.10	0.50	1.00	0.70	0.30	0.30	0.30	0.30	0.30	0.30	0.30
PV	0.40	0.40	0.40	0.40	0.40	0.50	1.00	0.70	0.70	0.70	0.70	0.70	0.70	0.70
DE	0.40	0.40	0.40	0.40	0.40	0.10	0.50	0.70	0.70	0.70	0.70	0.70	0.70	0.70
CA	0.05	0.05	0.05	0.05	0.05	0.80	0.70	0.50	0.90	0.90	0.80	0.30	0.30	0.30
DI	0.05	0.05	0.05	0.05	0.05	0.80	0.70	0.90	0.50	0.90	0.80	0.30	0.30	0.30
LM	0.05	0.05	0.05	0.05	0.05	0.80	0.70	0.90	0.90	0.50	1.00	0.50	0.30	0.30
CO	0.05	0.05	0.05	0.05	0.05	0.80	0.70	0.05	0.10	0.10	0.50	1.00	0.50	0.30
C2	0.05	0.05	0.05	0.05	0.05	0.80	0.70	0.05	0.05	0.10	0.10	0.50	1.00	0.50
EF	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.05	0.05	0.05	0.10	0.10	0.50	1.00
IM	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.50

Chain-level novelty coverage. Because Equation 6 scores candidate sequences by compatibility rather than by likelihood under a trained chain model, the inference procedure readily assigns a large δ to a sequence that is individually coherent but globally unseen. We quantify this via

$$n(S^*) = 1 - \max_{C \in \mathcal{C}} \text{LCS}(S^*, C) / |S^*|,$$

where $\mathcal{C}$ is a library of known campaigns (e.g., extracted from MITRE CTI reports). At *runtime* a reconstructed chain is flagged *coherent-but-novel* when its normalized score $\tilde{\delta}(S^*) \geq \tau_1$ (a confidence gate requiring no ground truth) and $n(S^*) \geq \tau_2$. For *evaluation*, where a per-scenario reference flow is available, we harden the confidence gate into a ground-truth accuracy gate—technique-LCS ≥ 0.5 against the reference (Section 7.4)—so the reported coherent-novel figure certifies that novel chains are *accurately reconstructed*, not merely far from known campaigns. The criterion supports both unknown campaign paths (high- n outputs) and partially unknown attacks (sequences whose δ is sustained by strong transition evidence despite locally ambiguous P_{ttp}).

Algorithm 2 Top- K Probabilistic Attack Chain Inference

Require: groups $\{b_1, \dots, b_T\}$, distributions $P_{\text{ttp}}(\cdot \mid b_t)$, beam K , transition-to-emission weight α , known-chain library $\mathcal{C}$

Ensure: best TTP sequence S^* , novelty score $n(S^*)$

```

1: for all  $t \in \text{Top}_K(P_{\text{ttp}}(\cdot \mid b_1))$  do
2:    $\delta[1][t] \leftarrow P_{\text{ttp}}(t \mid b_1)$ ;  $\psi[1][t] \leftarrow \perp$ 
3: end for
4: for  $s = 2$  to  $T$  do
5:   for all  $t_j \in \text{Top}_K(P_{\text{ttp}}(\cdot \mid b_s))$  do
6:      $\delta[s][t_j] \leftarrow 0$ 
7:     for all  $t_i \in \text{Top}_K(P_{\text{ttp}}(\cdot \mid b_{s-1}))$  do
8:        $v \leftarrow \delta[s-1][t_i] \cdot P_{\text{trans}}(t_i \rightarrow t_j)^\alpha \cdot P_{\text{ttp}}(t_j \mid b_s)$ 
9:       if  $v > \delta[s][t_j]$  then
10:         $\delta[s][t_j] \leftarrow v$ ;  $\psi[s][t_j] \leftarrow (s-1, t_i)$ 
11:       end if
12:     end for
13:   end for
14: end for
15:  $t_T^* \leftarrow \arg \max_t \delta[T][t]$ ;  $S^* \leftarrow \text{BACKTRACE}(\psi, t_T^*)$ 
16:  $n(S^*) \leftarrow 1 - \max_{C \in \mathcal{C}} \text{LCS}(S^*, C) / |S^*|$ 
17: return  $(S^*, n(S^*))$ 

```

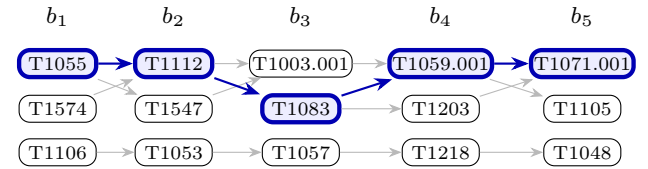


Fig. 3 Top- K inference lattice on the running example. Each column lists the Top-3 candidate techniques for a behavior group; gray arrows are scored candidate transitions and the highlighted path is the globally optimal chain S^* . At b_3 the path leaves the per-column emission top-1 (a_3) because the tactical-causal transition score favors the alternative — across the corpus, 21% of steps are resolved by transition rather than by emission alone. Under missing logs a dropped group simply disappears from the lattice and the surviving groups are linked directly, with the tactical matrix discounting the resulting cross-tactic jump.

6 Implementation

We implement SCOPE in Python 3.11 with `pandas` for log normalization, `networkx` for process-lineage construction, and `sentence-transformers` for the embedding stack; each design component of Section 5 maps to a self-contained module. Source code, anchor rules, and reproduction scripts are in our released artifact (Data availability).

Anchor rule set $\mathcal{R}$. The anchor rule set contains 46 templates derived from public detection rule libraries (Sigma [18], Atomic Red Team [30]) and re-curated for behavior-level recall rather than per-event precision. Each rule specifies a Sysmon/Windows event type, a small set of attribute predicates (e.g., command-line tokens, image-path patterns), and an associated technique *prior* that seeds—but does not bind—the candidate distribution (Section 5.3).

Embedding choice. SCOPE’s emission-side mapping must align long Sysmon-derived behavior descriptions with short MITRE technique abstracts (472 techniques across 14 tactics; v18.1), so we embed both with the general-purpose `all-mpnet-base-v2` [26, 27] (768-dim), whose broad pre-training generalizes across that length and style gap. The transition scores are non-neural — the tactical matrix M_{tac} and the causal set-overlap term — so no second embedding model is required.

Hybrid retrieval and reranking. Emission-side retrieval is a hybrid of dense cosine over mpnet embeddings and BM25 [31] (weight 0.15 over the Top-30 dense shortlist), with a TID-specific artifact-keyword *signature reranker* (weight 0.8 over the Top-10) acting as a tie-breaker among semantically close candidates. This dense-lexical-signature cascade preserves BM25’s token-level precision (specific registry hives, process names, named pipes) without sacrificing the dense channel’s ability to generalize across paraphrase.

LLM stage. The natural-language description $d_j = \text{Describe}(x_j)$ is generated by GEMINI-2.5-FLASH [32] via the hosted API. The prompt is a fixed template that consumes the structured feature vector x_j and constrains the output to a single-paragraph behavioral summary; no scoring or labeling decisions are delegated to the LLM, in line with the design separation discussed in Section 5.1. Every generated description is cached on disk and the cache ships with the artifact (Data availability), so every reported number reproduces without LLM API access.

Hyperparameters. The three design-relevant choices are: (i) an asymmetric temporal window (10s before, 30s after each anchor) reflecting that anchor-consequent evidence dominates anchor-antecedent

context; (ii) the transition-to-emission weight $\alpha = 0.5$ from a held-out sweep over $\{0.1, 0.3, 0.5, 0.7, 1.0\}$; and (iii) the tactical-causal fusion weights $w_{\text{tac}} = 0.5$, $w_{\text{cau}} = 0.2$ (renormalized over the retained axes). The complete numerical table (lineage radius, anchor budget, group cap, K , β , priors, self-loop penalty) is given in C, and the sweep-based sensitivity analysis in D shows that technique-LCS varies by at most ± 0.014 over the operational range of every hyperparameter — the chosen point sits on a plateau rather than a peak.

7 Evaluation

We organize the evaluation around five questions: **(Q1)** Does SCOPE reconstruct attack chains more accurately than event-, graph-, LLM-, and TTP-sequence baselines (Section 7.2)? **(Q2)** Does it remain robust as the fraction of missing observations grows (Section 7.3)? **(Q3)** Does it surface novel TTP and chain compositions outside the training/known-campaign distribution (Section 7.4)? **(Q4)** How much does each component contribute (Section 7.5)? **(Q5)** What is SCOPE’s cached offline throughput, and which components dominate runtime (Section 7.7)?

7.1 Experimental Setup

Datasets. We use 35 attack scenarios from the OTRF Security-Datasets (Mordor) corpus [33]: 33 *atomic* scenarios that exercise individual procedures across the eight ATT&CK tactics of COLLECTION, CREDENTIAL ACCESS, DEFENSE EVASION, DISCOVERY, EXECUTION, LATERAL MOVEMENT, PERSISTENCE, and PRIVILEGE ESCALATION, plus two *compound* multi-stage LSASS dump campaigns. Reference attack flows comprise 77 ground-truth steps in total. The grouping stage produces 779 behavior groups across the 35 scenarios under our per-rule anchor (20) and per-scenario group-size (80) caps; 251 are positively labeled as attack-relevant under a plausibility-based protocol (E), and the remaining 528 act as the benign-context distractor pool against which mapping and inference must remain selective. OTRF traces are short, atomic, and curated; we revisit this scope in Section 8.

Ground truth. We adopt a *plausibility-based* evaluation protocol: instead of a single canonical TTP sequence, each scenario is annotated with an *acceptable set* of valid attack flows derived from the published attack scripts and the corresponding ATT&CK Navigator layers. A predicted chain is matched against this set and credited for the most-aligned reference. This protocol explicitly accommodates the multi-realization

property of MITRE techniques and avoids penalizing semantically equivalent reformulations. Strict string-matching would treat predictions like **T1059** (parent) and **T1059.001** (sub-technique) of the same observed activity as misses, systematically underestimating any end-to-end method whose retrieval surfaces the more general technique.

Baselines. (i) **Sigma** [18], the de-facto rule-based event detector; (ii) **MAGIC** [34], a masked-graph-representation-learning detector applied to provenance subgraphs; (iii) **SHIELD** [8], the LLM-augmented graph-reasoning system; and (iv) **DeepAG** [23], originally a Transformer + bi-LSTM model over log-template sequences; we adapt the same architecture to operate over MITRE ATT&CK technique-id sequences (TID embedding in place of log-template embedding) so that its output is directly comparable to the chain-level metrics shared by all baselines, with training data drawn from the 53 layer files of our novelty library, exported from the MITRE ATT&CK v18.1 Campaigns catalog [35], and per-event Sigma classification as the test-time input stream. For (i)–(ii), which do not natively output a TTP sequence, we post-process detections into a chronologically ordered TTP list using each tool’s most confident per-event TTP attribution; the post-processing is described in F.

On baseline output asymmetry. DeepAG (next-token over a TID vocabulary) and MAGIC (per-node anomaly scores) do not natively emit ATT&CK labels; we pair each with Sigma as the downstream TID classifier (DeepAG refines Sigma’s per-event stream via its bi-LSTM; MAGIC suppresses Sigma alerts below the per-scenario median anomaly), whereas SHIELD and SCOPE label end-to-end. Thus the Sigma-vs-(DeepAG, MAGIC) gap measures reasoning layered *on top of* rule-based classification, while Sigma-vs-(SHIELD, SCOPE) measures fully end-to-end value. Each model’s native-task metric (DeepAG next-token accuracy; MAGIC entity AUC) is reported in Table 3.

Each adaptation puts the baseline on a *stronger-than-original* footing: DeepAG consumes Sigma’s already-classified TID stream rather than raw templates, MAGIC acts on Sigma-pre-filtered alerts, and SHIELD runs on the same GEMINI-2.5-FLASH as SCOPE’s description step rather than the original Qwen-2.5 32B; DeepAG also trains on the 53 CTI campaigns that double as SCOPE’s novelty library, an in-distribution advantage SCOPE does not share. The comparison is thus a best-effort adaptation, not a verbatim reproduction. Implementation details for the MAGIC re-implementation (PyTorch Geometric [36] port, benign-approximated training pool, per-process

Table 3 Native-task performance for the two baselines that do not natively output ATT&CK labels. DeepAG: next-TID Top- K accuracy on a 20% held-out slice of its 53 training campaigns. MAGIC: per-process anomaly ROC-AUC over the 31 of 35 scenarios that admit a valid positive/negative process split—the remaining 4 are skipped for lacking any positive process node (positive class = processes with events in any TP group).

Model	Native task	Score
DeepAG [23]	Next-TID Top-1 (held-out CTI)	0.18
DeepAG [23]	Next-TID Top-5 (held-out CTI)	0.44
MAGIC [34]	Per-process anomaly AUC (Mordor, $n=31$)	0.63

anomaly scoring) and the SHIELD reproduction (the four-module pipeline plus three documented adaptations for the OTRF setting) are deferred to G.

Metrics. *TTP-level*: Top-5 hit rate against the per-scenario acceptable set (a group hits if any of its five top candidates intersects the set). *Chain-level*: three normalized metrics over the recovered chain S^* vs. the closest reference — *technique-LCS* (longest common subsequence over technique IDs, reference-length-normalized), *tactic-LCS* (same on tactics), and *step coverage* (fraction of reference steps matched); strict and over-length-penalized variants are in Section 7.2 (Table 5). *Novelty coverage*: fraction of scenarios with $n(S^*) \geq \tau_2 = 0.30$ against the 53-campaign library [35] (the threshold is not load-bearing — shifting it moves all methods monotonically). *Robustness*: absolute technique-LCS drop from 0% to 50% uniform-random log loss, averaged over three random drop seeds (mean $\pm$ std; the other axes are seed-invariant once the LLM cache is populated).

7.2 Main Comparison (Q1)

Table 4 summarizes results across all 35 scenarios. SCOPE attains a Top-5 TTP hit rate of 0.84 and reconstructs attack chains with technique-LCS 0.68 and tactic-LCS 0.77, surpassing the strongest published baselines (DeepAG and MAGIC) by 29 points on technique-LCS. Event-level methods exhibit the steepest drop on chain-level metrics, reflecting their inability to reason across grouped behaviors. Graph- and LLM-based methods narrow the gap on TTP-level metrics but fail to maintain chain coherence when intermediate evidence is fragmentary. The ablation (Section 7.5) attributes most of the gap to behavior grouping and distributional mapping, both unavailable to event-, graph-, or per-event LLM baselines by construction.

Table 4 Main comparison on 35 OTRF scenarios. *Hit@5*: scenario-macro Top-5 plausibility over 251 attack-relevant groups; chain metrics macro-averaged (technique-LCS, tactic-LCS, step coverage), recall-normalized against acceptable references (strict variants in H). Best in **bold**. [†]SHIELD is evaluated here on short, sparse OTRF traces, outside its designed setting of long-duration, multi-host logs; see Section 7.2 for discussion.

Method	Family	Hit@5	tech-LCS	tac-LCS	step cov
Sigma [18]	Event	0.57	0.40	0.58	0.49
MAGIC [34]	Graph	0.57	0.39	0.52	0.48
SHIELD [†] [8]	LLM	0.23	0.13	0.13	0.14
DeepAG [23]	TTP-Seq	0.55	0.39	0.60	0.49
SCOPE (ours)	Beh. group	0.84	0.68	0.77	0.71

Strict and standard metrics across all meth-

ods. Because the recall-normalized chain-LCS credits sub-technique and contextual matches as positive evidence, we also report precision-oriented metrics under a *single uniform ruler applied identically to every method* (Table 5): strict-string technique-LCS (exact match, normalized by the longer chain), the over-length-penalized variant (family match, longer-chain normalized), per-step strict F_1 , and the false-chain rate (fraction of scenarios whose predicted chain shares no acceptable technique with the reference, with empty outputs counted as failures). Under exact-string matching every method scores low (≈ 0.10 – 0.11 technique-LCS), confirming that exact sub-technique recovery is hard for all systems and that the truth lies between the recall- and strict-normalized views. On the precision-aware composites, however, SCOPE leads: it attains the highest per-step strict F_1 (0.25) and the over-length-penalized LCS (0.15), and the lowest false-chain rate (5.7%, i.e. 2/35 scenarios), versus 22.9% for the rule/graph/sequence baselines and 74.3% for SHIELD, which emits an empty chain on 21 of 35 scenarios. This stems from SHIELD’s reliance on provenance-graph community analysis, which has little structure to exploit on OTRF’s short, sparse, often local-only traces; our reproduction runs it on the same LLM as SCOPE with an empty-graph fallback (G), so the gap reflects this sparsity rather than a weakened port — consistent with SHIELD’s own evaluation [8], which reports its lowest precision (0.08) on its most SCOPE-comparable dataset (the Windows, multi-host Public Arena) and reaches perfect precision only on the long-duration Linux/FreeBSD DARPA traces. Crucially, all baselines are re-implemented and scored with the *same protocol and ground truth* as SCOPE — we do not transcribe numbers from prior papers — so the comparison is internally consistent and not configured in SCOPE’s favor. Per-scenario strict breakdowns are provided in H.

Table 5 Strict/standard chain metrics under one uniform ruler over all methods (macro over 35 scenarios; empty chains = failures). strict-LCS: exact-string technique-LCS; len-pen: family-match LCS (both longer-chain normalized); F_1^{str} : per-step strict F_1 ; false: false-chain rate. Best in **bold**; strict-LCS is statistically indistinguishable across methods.

Method	strict-LCS	len-pen	F_1^{str}	false
Sigma [18]	0.099	0.108	0.229	22.9%
MAGIC [34]	0.102	0.122	0.232	22.9%
SHIELD [8]	0.114	0.129	0.109	74.3%
DeepAG [23]	0.101	0.111	0.235	22.9%
SCOPE (ours)	0.113	0.149	0.249	5.7%

7.3 Robustness to Incomplete Logs (Q2)

This experiment isolates the event-loss dimension of partial observability. We uniformly remove $p \in \{10, 25, 50\}\%$ of events per scenario and re-run all methods; Figure 4 reports technique-LCS at each drop rate.

The robustness curves separate into two regimes. Each drop rate is averaged over three random drop seeds for SCOPE and the three rule/graph/sequence baselines (shaded bands are ± 1 std); the spread is tight for every method ($\text{std} \leq 0.03$), so the curves reflect stable behavior rather than a single-mask artifact. SHIELD is shown for a single seed, as its degenerate empty-chain behavior is seed-insensitive. The four baselines (Sigma, MAGIC, DeepAG, SHIELD) degrade gently — 4 to 7 percentage points from 0% to 50% drop — because their underlying mechanism (per-event rule matching for Sigma/MAGIC/DeepAG, per-community LLM analysis for SHIELD) is largely stateless across events: dropping events independently scales the alert volume but preserves the alert-to-noise ratio. SCOPE starts much higher (0.68) and declines only gently to a 0.57–0.61 plateau (mean $\pm$ std over three seeds: 0.61 ± 0.03 , 0.59 ± 0.01 , 0.57 ± 0.03 at 10/25/50%): when a uniformly-random drop removes an anchor the corresponding behavior group is lost, but the surviving groups are still linked in observed order and the tactical matrix discounts the resulting cross-tactic jumps rather than breaking the chain, so chain quality stabilizes instead of collapsing. Crucially, SCOPE at 50% drop (0.57) still far exceeds the 0% chain-LCS of every baseline (max 0.40) — indeed its worst single seed at 50% (0.55) remains above every baseline’s intact-log value — so the absolute chain-quality ranking is preserved across the entire log-loss range and across seeds; the reconstruction advantage is not confined to complete observations. Coordinated adversarial deletion and enterprise noise are discussed in Section 8.

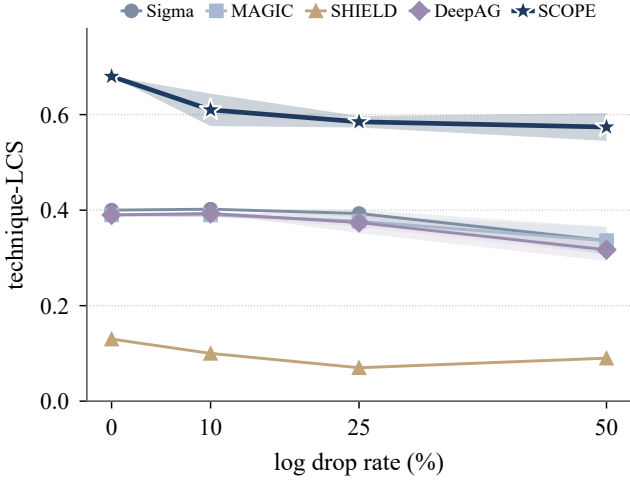


Fig. 4 Technique-LCS vs. random log-drop rate (mean over three random drop seeds; shaded bands are ± 1 std, SHIELD single seed). Baselines degrade gently but start much lower; SCOPE declines from 0.68 to a 0.57–0.61 plateau across 10–50% drop, its band remaining far above every baseline’s throughout.

7.4 Novelty Coverage (Q3)

Table 6 reports three metrics. *Rule-prior deviation* is a property check on the candidate distribution—the fraction of groups whose top-1 candidate differs from the anchor rule’s tagged TID; rule-bound baselines cannot deviate from their anchor TID by construction (marked †). *Chain-novel* alone is misleading because a chain with no overlap with the reference still counts as novel under $n(S^*) \geq 0.30$; we therefore also report *coherent-novel*, which couples novelty with ground-truth quality (technique-LCS ≥ 0.5 against the reference) and is the headline figure. SCOPE attains coherent-novel 0.83 (29/35) — 32 pp above the next-best baseline, compared to a 20-pp gap on raw chain-novel. The gap widens because rule-bound baselines often produce technically novel-looking sequences (short, sparse tactic chains register as novel against a campaign by reordering alone) whose chains do not clear the quality bar. We emphasize that *novelty* here refers to coherent recombinations of known ATT&CK techniques, not the discovery of new behavior classes.

7.5 Ablation Study (Q4)

Figure 5 reports technique-LCS when individual components are disabled. The single largest contributor is behavior grouping: replacing each group with its anchor event alone drops technique-LCS to 0.55 (−13 pp), and removing the shared-entity criterion (lineage only) drops it to 0.65. Distributional mapping

Table 6 Per-group rule-prior deviation and chain-level novelty. *Rule-prior dev.* (top-1 differs from the anchor prior at margin ≥ 0.02) is a property of the candidate distribution, not a discovery claim. †rule-bound methods cannot deviate from their anchor TID by construction. ‡SHIELD: 9 out-of-envelope TIDs over the 779-group denominator give 0.01. *Chain-novel* ($n(S^*) \geq 0.30$) ignores quality; *coherent-novel* adds a quality gate (technique-LCS ≥ 0.5) and is the figure to compare.

Method	Rule-prior deviation	Chain-novel	Coherent-novel
Sigma	—†	0.74	0.51
MAGIC	—†	0.69	0.43
SHIELD	0.01‡	0.00	0.00
DeepAG	—†	0.74	0.51
SCOPE	0.36	0.94	0.83

is the next largest — collapsing the lattice to top-1 gives 0.62 (−6 pp). Among the transition factors, tactical and causal each contribute ~ 3 –4 pp (0.64 and 0.65 respectively); transition compatibility is driven by tactical and causal structure. The fusion weight itself is non-trivial: $\alpha \rightarrow 0.1$ (emission-dominated) drops to 0.65, $\alpha \rightarrow 1.0$ (equal-weight) collapses to 0.60. Reverting the wildcard-IN refinement moves the score by only one point (0.68 $\rightarrow$ 0.67). The emission-confidence bypass is orthogonal to technique-LCS but its removal drops tactic-LCS from 0.77 to 0.74 and order accuracy from 0.60 to 0.57. Overall, no single component is sufficient and grouping, distributional mapping, and structural inference are largely complementary. Semantic information enters SCOPE through behavior descriptions and distributional TTP mapping—the −13pp grouping and −6pp top-1 collapse drops above—while the transition scores encode tactical order and causal artifact flow.

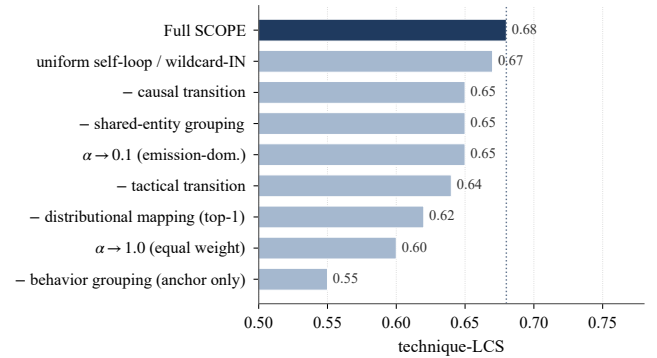


Fig. 5 Component ablation, technique-LCS (dotted line = Full SCOPE, 0.68). Behavior grouping is the largest contributor (−13 pp), then distributional mapping (−6), tactical (−4), and causal/shared-entity (−3 each).

Table 7 LLM-swap ablation: only description generation changes, all structural components fixed (macro over 35 scenarios). A comparable-tier model from another vendor yields near-identical chains; removing the LLM (deterministic template over the same features) isolates the semantic-abstraction role, not a specific frontier model.

Description source	tech-LCS	tac-LCS	step cov	order
GEMINI-2.5-FLASH (default)	0.683	0.771	0.707	0.600
gpt-4o-mini (other vendor)	0.660	0.767	0.700	0.567
Template (no LLM)	0.374	0.512	0.402	0.471

Frontier-model contribution (LLM swap).

To isolate how much of SCOPE’s performance is attributable to the specific frontier model rather than to the framework, we swap only the description-generation step — holding grouping, embedding, retrieval, and Viterbi inference fixed — and re-run end to end (Table 7). Replacing GEMINI-2.5-FLASH with a comparable-tier model from a different vendor (OpenAI gpt-4o-mini) changes technique-LCS by only 0.023 (0.683 $\rightarrow$ 0.660) and leaves tactic-LCS and step coverage essentially unchanged: the result is not Gemini-specific, and any capable LLM at this tier recovers the same chains. Removing the LLM entirely — replacing the description with a deterministic template rendered from the *same* structured features — drops technique-LCS to 0.37. This gap measures the value of the *semantic-abstraction role* the LLM plays (turning raw telemetry into technique-relevant natural language), not of any particular frontier model: SCOPE is robust to LLM choice while still benefiting from having one, and the structural inference retains substantial signal even with no LLM at all.

7.6 Case Study: *empire_mimikatz_logonpasswords*

We illustrate the cross-method gap on a canonical credential-dumping scenario in which an Empire-launched PowerShell process executes `Invoke-Mimikatz logonpasswords` against `lsass.exe`. The reference flow is T1059.001 (EXECUTION) $\rightarrow$ T1003.001 (CREDENTIAL ACCESS).

SCOPE recovers both reference techniques in their correct kill-chain order, yielding technique-LCS = 1.0 and step coverage = 1.0 under the recall-normalized metric of Section 7.1; the recovered chain itself contains 15 behavior groups in total, of which the two reference TTPs appear in order alongside contextual surrounding steps (the longer-chain-normalized strict variant is therefore 2/15, reported in H). Mean per-step similarity to the top-1 MITRE reference is $\tilde{\delta}(S^*) = 0.69$, and chain-level novelty against the 53-campaign library is $n(S^*) = 0.67$ (closest campaign: *Operation Wocao*,

C0014); the chain clears both the runtime confidence gate ($\tilde{\delta} = 0.69 \geq 0.5$) and—under ground truth—the accuracy gate (technique-LCS = 1.0 ≥ 0.5), so it is *coherent-novel* ($n \geq 0.30$) by both criteria.

The same scenario exposes each baseline’s failure mode. SHIELD’s LLM analyzer is drawn to the largest community (two `svchost.exe` instances repeatedly modifying a non-standard registry key) and mislabels the chain as T1547.001 + T1036.005, missing the Empire+Mimikatz pair entirely (Hit@5 = 0). Sigma and DeepAG fire on `rundll32/lsass` alerts and recover T1003.001 as a top candidate, but interleave it with NTLM-logon false positives that drag chain-LCS to 0.50; MAGIC, riding Sigma’s stream after anomaly-based filtering, mirrors Sigma’s outcome. SCOPE’s behavior-grouping step instead binds the Empire shell, the elevated launcher, and the `lsass`-touching credential read into a single coherent unit; the distributional mapping surfaces T1003.001 as the dominant candidate, and the tactical-causal transition score keeps T1059.001 and T1003.001 aligned in the lattice.

7.7 Efficiency (Q5)

On a single CPU-only workstation (Intel Core i7-12700F, 32 GB RAM, no GPU), SCOPE processes the full 35-scenario corpus (388,561 events) in 844 s — an aggregate throughput of 460 events/sec, per-scenario median 264 events/sec, peak resident memory 2.1 GB. With Gemini and the technique-IO table read from disk cache (the cached-deployment regime; cold-run LLM latency separation is left for follow-up work, Section 8), the cost is concentrated in the preprocessing stages: behavior grouping is the largest at 41%, followed by distributional mapping (26%) and feature extraction (21%), with event load/normalization (12%) and the deterministic Top- K Viterbi recurrence itself only 0.3%. Measured on the same corpus and the same cached-offline regime—excluding cold LLM generation latency for both systems—our SHIELD re-implementation runs at 77 events/sec, so SCOPE is roughly 6 $\times$ faster end-to-end while producing a structured, scored chain rather than a free-text community explanation.

7.8 Learning-free vs. Supervised Emission (Q6)

SCOPE maps each behavior group to techniques by *learning-free* nearest-neighbor retrieval against the full MITRE ATT&CK catalogue, rather than by a supervised classifier trained on labeled groups. To justify this

Table 8 Learning-free (SCOPE) vs. supervised emission under leave-one-scenario-out CV, same per-group plausibility metrics over 251 labeled groups. *seen-only*: optimistic upper bound evaluating only held-out groups whose technique also appears in training. Best in **bold**.

Emission method	strict H@1	plaus. H@1	plaus. H@5
SCOPE (learning-free)	0.38	0.66	0.86
Linear SVM (TF-IDF), LOSO	0.33	0.49	0.87
CatBoost (features), LOSO	0.22	0.36	0.82
Linear SVM, seen-only	0.46	0.60	0.91
CatBoost, seen-only	0.30	0.42	0.89

choice, we train supervised emission baselines on the 251 labeled groups and evaluate them under leave-one-scenario-out (LOSO) cross-validation — the realistic test of generalizing to an unseen scenario — against the identical plausibility metrics used for SCOPE’s mapper (Table 8). We test both a standard text classifier (TF-IDF over the same feature text $\rightarrow$ linear SVM) and a strong tabular model (engineered features $\rightarrow$ gradient-boosted trees, CatBoost). On top-1 accuracy, SCOPE’s learning-free emission (plausible H@1 = 0.66, micro-averaged over the 251 groups to match the per-group supervised protocol; the scenario-macro value is 0.62, Table 9) outperforms every supervised variant, including the *optimistic* seen-only upper bound that discards held-out groups whose technique never appears in training (best supervised 0.60). The stronger model does not help — CatBoost trails the simpler SVM — indicating the bottleneck is data, not model capacity. The cause is structural: of the 30 distinct techniques in the corpus, 24 occur in only one scenario, so under LOSO roughly a quarter of test groups carry a technique the classifier has never seen and therefore cannot predict, whereas SCOPE retrieves against the entire 472-technique catalogue without any training. Supervised emission is thus viable only with abundant, well-balanced labeled attack data; the learning-free design generalizes to the long tail of rare and unseen techniques that dominates realistic telemetry.

7.9 Per-Submodule Diagnostics (Q7)

To localize where errors arise, we evaluate each stage in isolation against the manually labeled groups (Table 9). The *grouping* stage has purity 0.47 (fraction of produced groups that are attack-relevant) and completeness 0.42 (fraction of reference techniques covered by at least one correctly-labeled group), reflecting the deliberate recall-first anchor design that admits benign-context distractors for downstream filtering. The anchor stage itself has broad reach — at least one anchor fires on a real attack step in 33/35 scenarios, and anchor rules reach 64% of reference step-types (macro) —

Table 9 Per-submodule diagnostics (macro over 35 scenarios). Grouping: purity = attack-relevant fraction, completeness = reference techniques covered. Mapping (TP groups only): plausibility H@1/H@5, MRR. Hallucination proxy: fraction of FP groups given a plausible top-1.

Stage	Metric	Value
Grouping	Purity	0.47
Grouping	Completeness	0.42
Mapping (TP)	Plausible H@1	0.62
Mapping (TP)	Plausible H@5	0.84
Mapping (TP)	MRR	0.70
Mapping	Hallucination proxy (FP $\rightarrow$ plausible)	0.20

so the lower labelled-completeness reflects downstream labelling strictness rather than the anchor net missing the activity. The *mapping* stage, evaluated only on true-positive groups so that its competence is not confounded by grouping false positives, attains plausible H@1 = 0.62, H@5 = 0.84, and MRR = 0.70 (macro-averaged over scenarios; the per-group micro averages of 0.66/0.86 are the ones compared against the supervised baselines in Section 7.8). As a hallucination proxy, 20.5% of false-positive (non-attack) groups still receive a top-1 technique inside the scenario’s acceptable set — the mapper occasionally assigns a plausible-looking technique to benign context — which the distributional representation absorbs by keeping a candidate distribution instead of committing to a single label, and which the structural transition scoring further down-weights when a benign group does not fit a coherent chain. These diagnostics confirm that grouping recall and the structural inference stage, rather than the mapper’s per-group accuracy, are the dominant levers, consistent with the ablation in Section 7.5.

8 Discussion

Anchors: triage, not detection. Anchors decide only *where* deeper analysis should focus; the technique decision is made separately and semantically against the entire ATT&CK technique set (Section 5.3), and 36% of behavior groups end on a technique different from their anchor’s prior (Table 6, rule-prior dev.). Selecting suspicious regions for closer inspection is the standard, unavoidable first step of any analysis pipeline; SCOPE’s contribution is not that it uses such a net but that it *decouples the net from the technique decision*, so a coarse anchor triggering merely in the right vicinity suffices and rule-flagged activity is re-interpreted rather than trusted. Recall is thus deliberately preserved at the anchor stage while precision is refined downstream: the distributional mapping avoids an early hard commit-

ment, and the final technique is decided by semantic similarity against the entire ATT&CK set, so a group that matches no technique well carries a low similarity and confidence margin exposed per group for triage. The one genuine bound is completeness — a behavior invisible to every rule yields *a missed group, not a wrong chain*, a limit shared by all detection-based analysis and one that stays localized, leaving the rest of the recovered chain intact. SCOPE is therefore not anchor-free discovery: it reconstructs coherent chains once sparse suspicious evidence exists, rather than replacing low-level telemetry triggers.

LLM dependency. The description-generation step is the only path for LLM error to enter the pipeline; inference itself is deterministic and never queries the LLM. In manual review, most omissions of non-essential details (e.g., a specific registry hive) did not flip the retrieval top-1 because the joint feature vector x_j and the BM25 channel anchor the match to actually-observed tokens. *Counterfactual* hallucinations that introduce non-existent artifacts are more concerning, but were observed infrequently and rarely flipped the top-1 for the same reason; we treat this as qualitative evidence rather than a measured hallucination rate. Constrained decoding with a stricter template would tighten this further at modest throughput cost.

Threats to validity. We use OTRF because it gives every method the same controlled evaluation surface: short, atomic traces with a published reference flow per scenario, so chain-LCS is well-defined against a known ground truth. Public corpora are in fact scarce in this respect: most provenance datasets provide node- or graph-level malicious/benign labels rather than ordered technique sequences, and CTI corpora carry no host telemetry, so OTRF is among the few that pair raw endpoint logs with the per-scenario *ordered technique* ground truth that chain-level evaluation requires. Production traffic exhibits longer benign tails and richer concurrency, both of which raise per-step ambiguity — precisely the regime that SCOPE’s distributional mapping and tactical-causal transition score were built for. The 50% random-drop result in Section 7.3 (technique-LCS 0.57, still well above every baseline’s 0% value) is the closest direct evidence we have that this design tolerates real-world fragmentation, but a full enterprise evaluation remains outside this paper’s scope.

Our drop-based study models uniform event loss but not coordinated tampering of high-value records, which would warrant adversary-aware perturbations [2]; an adaptive adversary specifically deleting SCOPE’s anchor events would suppress the corresponding behavior groups but cannot, by construction,

inject a spurious group, so the failure mode collapses to the anchor-recall localization above. More broadly, structured drop modes each exercise a different design feature: event-type-targeted deletion is partially absorbed by the shared-entity criterion (a registry write missing from one view often survives in process lineage), contiguous time-window bursts remove consecutive groups, which surviving-group linking and tactical-jump discounting absorb, and anchor-targeted deletion reduces to the same localization. A clean benign-only false-alarm study, calibrated probability evaluation of P_{ttp} (e.g., expected calibration error), and cold-run LLM latency separation are likewise informative follow-up measurements we deliberately did not include in this round. Reported numbers for SHIELD and MAGIC depend on our re-implementations on Windows-only telemetry; absolute comparisons should be read together with the family-level positioning in Section 3.

Generalization beyond the training distribution. SCOPE’s transition score is built from domain priors rather than from a chain model trained on a fixed campaign distribution (Section 5.4). When the attack vocabulary shifts (a new ATT&CK technique, a new tool family manifesting through known behaviors, or shifted tactic frequencies), only the anchor rule library and the MITRE reference index need refreshing, both of which are public. The framework is OS-agnostic at the implementation level; a full Linux/macOS evaluation is left for future work.

Streaming readiness. Behavior grouping is naturally time-windowed and the Top- K Viterbi recurrence operates on already-emitted groups in chronological order, so the throughput of Section 7.7 is indicative of the non-LLM incremental cost of a streaming deployment; a complete streaming-mode evaluation is left to future work.

9 Conclusion

This paper reframes attack-chain reconstruction as uncertainty-aware reasoning under partial observability, rather than as classification-style mapping over complete logs. SCOPE realizes this view through three coupled mechanisms — behavior-level abstraction, distributional TTP mapping, and a Top- K Viterbi lattice — that together resolve the tension between semantic ambiguity (which forbids early commitment to a single TTP) and structural consistency (which demands a single coherent chain) by deferring commitment until sequence inference time.

Across 35 OTRF scenarios, SCOPE attains a Top-5 hit rate of 0.84 and technique-LCS of 0.68, surpassing

four baseline families and holding at 0.57 under 50% random log drop — above every baseline’s intact-log score — while producing a false chain on only 2 of 35 scenarios (5.7%), versus 22.9% for the rule- and graph-based baselines. Beyond the empirical gap, our results suggest that robustness to incomplete and noisy telemetry cannot be obtained from better per-event classification alone — it requires explicit modelling of uncertainty at both the semantic and structural levels. Cross-host correlation and long-horizon enterprise validation remain important next steps.

Declarations

Funding This work was supported by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2026-25509873, Developing a Simulator for Cyber Defense Maturity Assessment and Autonomous Security) (50%), and by the Korea Internet & Security Agency (KISA) grant funded by the Korean government (PIPC) (No. RS-2025-02222626, Development of Technology for Assessing Privacy Vulnerabilities and Diagnosing Protection Levels of Generative AI Models) (50%).

Competing interests The authors have no competing interests to declare that are relevant to the content of this article.

Ethics approval This work uses only the publicly available OTRF Security-Datasets (Mordor) corpus, captured in a controlled laboratory environment; it involves no human participants, animals, or personally identifiable information, so ethics-committee approval and informed consent are not applicable. Ethical considerations relating to the defensive nature of the work are discussed in Appendix I.

Author contributions J.L. conceived and designed the study, developed and implemented the framework, carried out most of the experiments and the analysis, prepared the figures, and wrote the abstract and the bulk of the manuscript. J.O. assisted with the attack-chain reconstruction experiments. S.L. contributed substantially to writing the front sections of the paper, from the Introduction through the Related Work (Sections 1–3). T.L. supervised and managed the project, critically reviewed the manuscript, and acquired the funding. All authors read and approved the final manuscript.

Data availability The source code, anchor rules, reference attack-flow annotations, and the on-disk LLM-description cache required to reproduce every reported number are released as a public artifact at <https://github.com/jylee092/SCOPE>, together with the four reproduced baselines and the reproduction harness. The OTRF Security-Datasets (Mordor) corpus used for evaluation is publicly available under its original license.

Table 10 Tactical transition types encoded in M_{tac} . Weights are pre-defined and reflect kill-chain semantics rather than empirical frequencies.

Transition type	Wt.	Example
Primary flow (canonical adjacent)	1.00	Execution → Persistence
Post-exploit loop (direct cycle)	0.90	Discovery → Lateral Mov.
Post-exploit loop (convergent)	0.80	Lateral Mov. → Collection
Wildcard-IN (to privilege)	0.80	any → Privilege Esc.
Wildcard-IN (to defense evasion)	0.70	any → Defense Evasion
Wildcard-OUT (forward)	0.70	Defense Evasion → Collection
Self-loop	0.50	C2 → C2
Forward skip (2–3 stages)	0.50	Initial Access → Persistence
Wildcard-OUT (reverse)	0.40	Defense Evasion → Initial Access
Far skip (4+ stages)	0.30	Recon → Execution
Backward, near (1–2 stages)	0.10	Execution → Initial Access
Backward, far (3+ stages)	0.05	Collection → Initial Access
Forbidden (post-impact re-set)	0.02	Impact → Reconnaissance

A Running Example Walkthrough

We thread a fragment of the *msf_record_mic* OTRF scenario [33] through Section 5; its chain spans PERSISTENCE → PRIVILEGE ESCALATION → DEFENSE EVASION → EXECUTION → COMMAND-AND-CONTROL, and we additionally assume an intermediate DISCOVERY step is unobserved (e.g., audit-buffer tampering or a restrictive retention policy) [1, 17]. Anchors include the PowerShell invocation, the *fodhelper.exe* UAC bypass, the audit-log clearing, the *rundll132* Meterpreter execution, and the HTTP beacon; process lineage plus the shared-entity criterion (e.g., the *fodhelper.exe* registry write to the elevated process that consumes the key, and the dropped payload to its later *rundll132* execution) bind these into the behavior groups feeding the lattice of Figure 3.

B Tactical Transition Matrix

Table 10 lists every named transition type encoded in M_{tac} (Section 5.4), with its weight and a representative example. Weights are pre-defined from kill-chain semantics, not learned from corpus frequencies, which is what gives SCOPE its training-free generalization to unseen TTP compositions (Section 8).

C Hyperparameters

Table 11 summarizes every numerical hyperparameter referenced in Sections 5–7. All values are deterministic and were either taken directly from a cited prior work (e.g., SHIELD’s LOF $k=20$, MAGIC’s mask rate 0.5) or fixed via the held-out parameter sweeps logged at `Final_Code/run_log_*.txt`.

D Hyperparameter Sensitivity

To verify the operating point is not knife-edge tuning, we swept every hyperparameter introduced with a specific value, fixing the rest at their C defaults and evaluating on the same 251 groups across 35 scenarios with the on-disk LLM cache (fully deterministic; per-variant numbers at `Final_Code/output/v*_sweep_results.json`). Figure 6 shows the two widest-varying axes: the transition-to-emission weight α and the emission-confidence bypass threshold.

Other axes. Three further sweeps (wildcard-IN soft-tuning, decoder-mode variants, and the joint $\alpha \times$ bypass grid) produced even smaller variation and are omitted for space; the worst grid combination still exceeds every baseline’s 0%-drop chain-LCS. We additionally stress the tactical matrix directly: jittering *every* M_{tac} weight by $\pm 10/20/30\%$ (20 random trials each) moves technique-LCS from 0.683 to means of 0.672/0.663/0.656, and the worst of the 60 perturbed configurations (0.617) still exceeds every baseline’s intact-log chain-LCS (0.40) — the reconstruction is not knife-edge tuned to the specific hand-set matrix values.

Two takeaways: (i) any operating point inside the shaded ranges of Figure 6 changes technique-LCS by at most 1.4 pp and SCOPE still outperforms every baseline; (ii) the only failure modes are the degenerate ones the formulation predicts ($\alpha = 1.0$ disables the emission anchor; a bypass threshold below the similarity floor fires on noise), both excluded by construction.

E Plausibility-Based Annotation Protocol

The 251 attack-relevant behavior groups are labeled under a plausibility protocol that accepts any technique consistent with the scenario’s published attack flow rather than a single canonical TTP; it is fully scriptable and is what *Hit@5* and the chain-level metrics in Section 7 score against.

Reference flow. For each of the 35 OTRF scenarios we curate an ordered reference flow from the Mordor script and its ATT&CK Navigator layer, each step a tuple (tid,tactic,alts,note) where *alts* lists alternative realizations (e.g., for an Empire *logonpasswords* command both T1003.001 and the parent T1003); the full set is open-sourced at `experiments/attack_flows.py`.

Acceptable set and matching. Per scenario the acceptable set is the union of every step’s primary TID and its alternatives; a predicted TID matches if it equals an acceptable TID or shares its parent (parent–child collapse), absorbing the ATT&CK sub-technique ambiguity (e.g., T1059 vs. T1059.001) that dominates strict-string error on OTRF. The collapse rule derives solely from ATT&CK’s official parent/sub-technique numbering—not from any method’s output—and, like the acceptable set, is applied identically to every method’s predictions, so the protocol is agnostic to whether a method emits techniques at the parent or sub-technique granularity.

Per-group ground truth. A group is *TP* if any of its events’ anchors matches an acceptable TID by family (sub-categorized *strong/weak* by whether the anchor or only supporting events carry the attack tag; both count as TP); the remaining 528 of 779 groups form the FP/distractor pool. Labels are produced by `experiments/auto_label_gt.py` and scored by `run_eval_plausible.py` (Hit@5) and `chain_align.py` (chain metrics).

Table 11 Hyperparameters used throughout this work.

Stage	Symbol / config key	Value
<i>Behavior grouping (Section 5.2)</i>		
window before anchor	GROUPING_BEFORE_SEC	10 s
window after anchor	GROUPING_AFTER_SEC	30 s
lineage hop up	GROUPING_HOP_UP	2
lineage hop down	GROUPING_HOP_DOWN	3
anchor budget per rule	GROUPING_MAX_ANCHORS_PER_RULE	20
group cap per scenario	MAX_GROUPS_PER_SCENARIO	80
group merge overlap threshold	MERGE_OVERLAP_THRESHOLD	0.5
<i>Distributional TTP mapping (Section 5.3)</i>		
description encoder	EMBED_MODEL_NAME	all-mpnet-base-v2
top-K candidates per group	TOP_K	5
softmax temperature	β	10.0
BM25 hybrid weight	BM25_WEIGHT	0.15
BM25 rerank width	BM25_RERANK_WIDTH	30
TID-signature rerank weight	SIGNATURE_WEIGHT	0.8
TID-signature rerank width	SIGNATURE_RERANK_WIDTH	10
rule-TID prior boost	RULE_TID_PRIOR	1.15
rule-tactic prior boost	RULE_TACTIC_PRIOR	1.05
emission-confidence bypass thr.	EMISSION_BYPASS_SIM_THRESHOLD	0.75
LLM (description only)	GEMINI_MODEL	gemini-2.5-flash
<i>Two-axis transition (Section 5.4)</i>		
fusion weight, tactical	w_{tac}	0.5
fusion weight, causal	w_{cau}	0.2
<i>Probabilistic sequence inference (Section 5.5)</i>		
beam width	K	5
transition-to-emission weight	α	0.5
self-loop TID penalty	SELF_LOOP_TID_PENALTY	1.0
hard tactic-mismatch filter	VITERBI_HARD_TACTIC_FILTER	on
<i>Novelty thresholds (Section 7.4)</i>		
Rule-prior deviation margin	μ	0.02
chain-novel threshold	τ_2	0.30
<i>Baselines (Section 7.1)</i>		
SHIELD LOF neighbours	k (LOF)	20
SHIELD LOF contamination	τ	0.10
SHIELD CoT confidence threshold	δ	0.80
MAGIC mask rate	—	0.50
MAGIC GAT layers / hidden / heads	—	3 / 64 / 4
MAGIC KNN neighbours	k	10
DeepAG window length	h	10
DeepAG LSTM hidden	a	64
DeepAG override threshold	—	0.50

F Per-Event TTP Attribution Post-Processing

Sigma, MAGIC, and DeepAG do not natively emit a single ordered TTP sequence; we post-process each into a chronological list scored by the same chain-LCS metrics as SHIELD and SCOPE.

Sigma. Per event we evaluate the 2,140 SigmaHQ Windows rules, rank firing rules by (`level`, `status`, `selection-count`, `rule-id`), union their `attack.tNNNN` tags in priority order, and keep the top-5; the scenario chain is the time-ordered top-1 TIDs with consecutive duplicates collapsed.

MAGIC. A masked-graph autoencoder yields per-process anomaly scores ($k=10$). MAGIC produces no TIDs, so we reuse Sigma’s per-event Top-5 and keep only alerts whose source process scores at or above the per-scenario

median anomaly — using MAGIC as a structural filter (Section 7.1).

DeepAG. The Transformer+bi-LSTM consumes the chronological Sigma top-1 TID stream ($h=10$) and, where its forward top-1 probability ≥ 0.5 , overrides Sigma’s top-1; the top-5 otherwise preserves Sigma’s coverage. All three chain outputs are time-ordered (technique-id, tactic) pairs with consecutive duplicates collapsed.

G Baseline Implementation Details

This appendix documents the MAGIC and SHIELD reproductions; Section 7.1 summarizes the high-level adaptations.

MAGIC. We re-implement the masked-graph autoencoder of Jia et al. [34] in PyTorch Geometric, training it

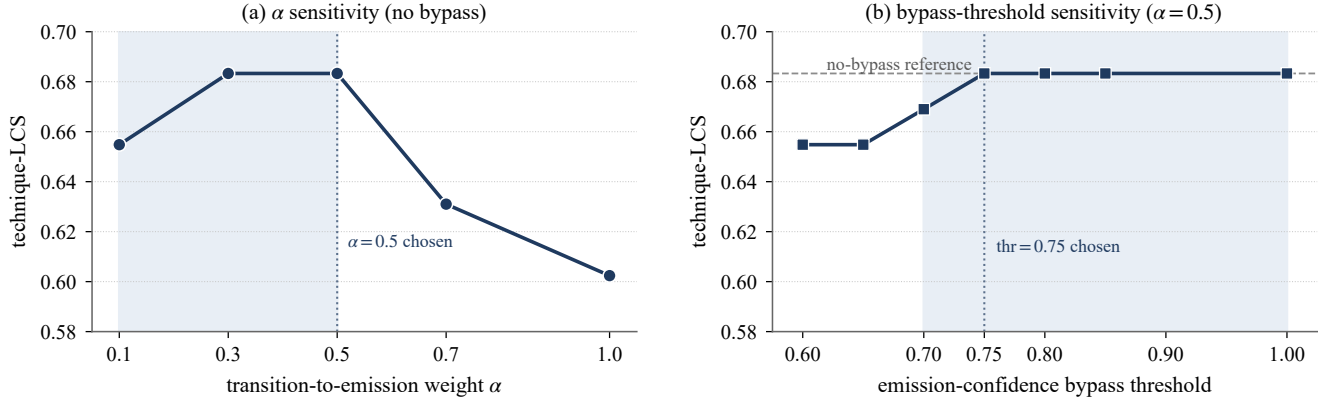


Fig. 6 Hyperparameter sensitivity. **(a)** Transition-to-emission weight α (with bypass disabled). Across the operational range $\alpha \in [0.1, 0.5]$ (shaded) technique-LCS varies by ± 0.014 ; larger weights degrade steadily ($\alpha = 0.7 \rightarrow 0.631$, $\alpha = 1.0 \rightarrow 0.602$ where transition becomes equal-weight to emission). **(b)** Emission-confidence bypass threshold (with $\alpha = 0.5$ fixed). The dashed line marks the no-bypass reference (Viterbi only). Technique-LCS plateaus at 0.6833 for thresholds ≥ 0.75 , matching the no-bypass reference (the bypass mainly improves tactic-LCS and ordering, Section 7.5); thresholds ≤ 0.70 dip to 0.655–0.669. The chosen operating points ($\alpha = 0.5$, thr = 0.75) sit on the plateau in both panels.

self-supervised on benign-approximated Mordor sub-graphs (LOF-flagged anomalous nodes removed); node features are a type one-hot, a 16-dim identity hash, and type-specific flags. At test time per-process anomaly is the mean distance to the $k=10$ nearest benign neighbours, used to suppress Sigma alerts below the per-scenario median — preserving Sigma’s coverage while trimming low-anomaly noise (MAGIC *identifies* entities rather than labelling techniques).

H Strict-Metric Supplement

Table 12 Strict-metric supplement for SCOPE (35 scenarios). Strict variants forbid parent/sub-technique collapse; *plausible* variants apply parent collapse but no acceptable-set lenience. *False-chain rate* is the fraction of scenarios whose recovered chain has no overlap with any acceptable reference TID.

Metric (strict supplement)	Value
Strict-TID LCS (no parent collapse)	0.11
Per-step Precision (strict)	0.24
Per-step Recall (strict)	0.44
Per-step F1 (strict)	0.25
Per-step Precision (parent-collapsed)	0.29
Per-step Recall (parent-collapsed)	0.67
Per-step F1 (parent-collapsed)	0.33
False-chain rate (no acceptable TID in recovered chain)	0.057

SHIELD. We reproduce the four-module pipeline of Gandhi et al. [8] (LOF $k=20$, contamination 0.1; provenance graph with suspicious-tag propagation; Louvain communities; three-stage chain-of-thought analyzer) with four adaptations: (a) the LLM is GEMINI-2.5-FLASH (matching SCOPE’s description model, so the LLM is not a confound) rather than the original Qwen-2.5 32B; (b) when a scenario has no external socket (many local-only OTRF traces, e.g. `ntdsutil`), tag-propagation seeds fall back to the LOF-flagged processes rather than yielding an empty graph; (c) each short OTRF trace is treated as one time window, skipping the temporal correlation engine; and (d) each community is analyzed at a confidence of 0.7 and *all* analyzed communities are retained, rather than gating at SHIELD’s alert threshold $\delta=0.8$, so as to recover as much attack signal as possible — a relaxation that favors SHIELD’s recall. Stage-3 per-community output is expanded into the chronological prediction sequence rather than collapsed, preserving SHIELD’s LLM chain reasoning.

Table 12 gives SCOPE’s per-scenario strict diagnostics, complementing the cross-method strict comparison of Section 7.2 (Table 5): strict-TID LCS (no parent collapse), per-step P/R/F1 (strict and parent-collapsed), and the false-chain rate. Strict figures are conservatively low because OTRF references are parent-level while SCOPE retrieves sub-techniques; the false-chain rate is a benchmark-level proxy (not a benign-only operational false-positive rate) at 0.057 (2/35).

I Ethical Considerations

This work uses the OTRF Security-Datasets (Mordor) corpus, a public, license-compatible release of attacker activity captured in a controlled lab environment from publicly documented adversary tooling (Empire, Mimikatz, Metasploit,

Covenant, PurpleSharp). No new exploits, malware, or attack capabilities are developed or released; SCOPE is a defensive analysis framework whose inputs are post-incident endpoint logs and whose outputs are technique-id sequences and attack-chain reconstructions. The reproduced baselines (Sigma, MAGIC, SHIELD, DeepAG) are likewise defensive systems. We do not study, target, or weaken any production deployment, and our experiments are run entirely on the public OTRF lab data; no real-user personally identifiable information is involved. We see no plausible misuse pathway that would not already be available via the underlying public datasets and prior published methods.

References

1. R. Sekar, H. Kimm, R. Aich, in *2024 IEEE Symposium on Security and Privacy (SP)* (IEEE, 2024), pp. 3571–3589
2. C. Vanini, J. Gruber, C. Hargreaves, Z. Benenson, F. Freiling, F. Breiting, in *Proceedings of the Digital Forensics Doctoral Symposium* (2025), pp. 1–8
3. J. Zeng, Z.L. Chua, Y. Chen, K. Ji, Z. Liang, J. Mao, in *NDSS* (2021)
4. B. Zhang, Y. Gao, C. Yu, B. Kuang, Z. Zhang, H. Kim, A. Fu, in *34th USENIX Security Symposium (USENIX Security 25)* (2025), pp. 607–624
5. J. Zhou, S. Ying, S. Wang, D. Zhao, J. Xiang, K. Liang, P. Liu, *IEEE Transactions on Dependable and Secure Computing* (2025)
6. W.U. Hassan, D. Li, K. Jee, X. Yu, K. Zou, D. Wang, Z. Chen, Z. Li, J. Rhee, J. Gui, et al., in *Proceedings of the 36th Annual Computer Security Applications Conference* (2020), pp. 165–178
7. J. Li, R. Zhang, J. Liu, *Electronics* **13**(1), 100 (2024)
8. P.A. Gandhi, P.N. Wudali, Y. Amaru, Y. Elovici, A. Shabtai, arXiv preprint arXiv:2502.02342 (2025)
9. W. Cheng, T. Zhu, S. Jing, J.P. Mei, M. Ma, J. Jin, Z. Weng, arXiv preprint arXiv:2503.03108 (2025)
10. W. Guo, J. Xue, Z. Liu, W. Han, J. Hu, *The Journal of Supercomputing* **81**(9), 1049 (2025)
11. T. Li, Y. Liu, Y. Liu, Y. Xiao, N.A. Nguyen, *Computers & Security* **97**, 101974 (2020)
12. S. Zhang, X. Xue, X. Su, *Electronics* **14**(2), 257 (2025)
13. Z. Li, L. Wang, Z. Wang, X. Shen, H. Xu, Y. Chen, S. Ji, *IEEE Transactions on Dependable and Secure Computing* (2025)
14. S. Benabderrahmane, P. Valtchev, J. Cheney, T. Rahman, in *2025 13th International Symposium on Digital Forensics and Security (ISDFS)* (IEEE, 2025), pp. 1–6
15. P.N. Wudali, M. Kravchik, E. Malul, P.A. Gandhi, Y. Elovici, A. Shabtai, arXiv preprint arXiv:2502.02337 (2025)
16. B.E. Strom, A. Applebaum, D.P. Miller, K.C. Nickels, A.G. Pennington, C.B. Thomas, MITRE ATT&CK: Design and philosophy. Tech. Rep. MP180360R1, The MITRE Corporation (2018)
17. W.U. Hassan, A. Bates, D. Marino, in *2020 IEEE symposium on security and privacy (SP)* (IEEE, 2020), pp. 1172–1189
18. F. Roth, T. Patzke. Sigma: Generic signature format for SIEM systems (2017). URL <https://github.com/SigmaHQ/sigma>. Presented at Hack.lu 2017
19. X. Qiu, M. Lv, T. Zhu, Y. Zhang, T. Chen, *Computers & Security* p. 104907 (2026)
20. A. Aly, E. Mansour, A. Youssef, in *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security* (2025), pp. 261–275
21. Y. Fu, Z. Ding, S. Zhang, D. Cao, Y. Wang, in *2025 IEEE 24th International Conference on Trust, Security and Privacy in Computing and Communications (Trust-Com)* (IEEE, 2025), pp. 524–531
22. S. Choi, J.H. Yun, B.G. Min, in *Proceedings of the 14th Cyber Security Experimentation and Test Workshop* (2021), pp. 41–48
23. T. Li, Y. Jiang, C. Lin, M.S. Obaidat, Y. Shen, J. Ma, *IEEE Transactions on Dependable and Secure Computing* **20**(1), 740 (2022)
24. Z. Li, J. Zeng, Y. Chen, Z. Liang, in *European symposium on research in computer security* (Springer, 2022), pp. 589–609
25. S. Akhtar, S. Khan, S. Parkinson, arXiv preprint arXiv:2502.00677 (2025)
26. Sentence-Transformers. all-mpnet-base-v2: A general-purpose sentence-embedding model. Hugging Face Hub model card (2021). URL <https://huggingface.co/sentence-transformers/all-mpnet-base-v2>
27. N. Reimers, I. Gurevych, in *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)* (2019), pp. 3982–3992
28. E.M. Hutchins, M.J. Cloppert, R.M. Amin, *Leading Issues in Information Warfare & Security Research* **1**(1), 80 (2011)
29. M. Ferrer-Oliva, J.A. Medina-Merodio, J.J. Martínez-Herraiz, C. Cilleruelo-Rodríguez, *Sensors* **25**(23), 7124 (2025)
30. Red Canary. Atomic Red Team: An open-source library of detection tests mapped to MITRE ATT&CK. GitHub repository (2024). URL <https://github.com/redcanaryco/atomic-red-team>
31. S. Robertson, H. Zaragoza, *The probabilistic relevance framework: BM25 and beyond*, vol. 4 (Now Publishers Inc, 2009)
32. Gemini Team, arXiv preprint arXiv:2507.06261 (2025)
33. R. Rodriguez, J. Rodriguez. OTRF security datasets (2020). URL <https://github.com/OTRF/Security-Datasets>. Open Threat Research Forge
34. Z. Jia, Y. Xiong, Y. Nan, Y. Zhang, J. Zhao, M. Wen, in *33rd USENIX Security Symposium (USENIX Security 24)* (2024), pp. 5197–5214
35. The MITRE Corporation. MITRE ATT&CK campaigns. ATT&CK knowledge base, v18.1 (2025). URL <https://attack.mitre.org/campaigns/>. Catalog of documented threat-actor campaigns; layer files exported via ATT&CK Navigator
36. M. Fey, J.E. Lenssen, arXiv preprint arXiv:1903.02428 (2019)