

LASER: Defeasible Certificate Search for Long-Range Protocol Dependencies

Zihan Tang, Huibiao Zhu*

Shanghai Key Laboratory of Trustworthy Computing, East China Normal University, Shanghai, China
Email: 51285902038@stu.ecnu.edu.cn, hbzhu@sei.ecnu.edu.cn

Abstract

Protocol specifications rarely justify executable behavior in one place: a transition may inherit its source state from a remote control block, name its events only by number, and remain subject to a later correction. Recovering the nearby clause is therefore insufficient; the system must assemble a source-linked certificate and determine whether any stronger rule applies. We introduce LASER, a proof-carrying search method for sparse, defeasible protocol evidence. LASER represents missing support and unresolved corrections as explicit obligations, searches typed evidence routes, and lets an occurrence-aware symbolic checker alone decide whether a certificate warrants, refutes, or leaves a behavior unresolved. On 91 BehaviorCheck families under a matched 2K budget, LASER improves Joint Edge+Proof over the strongest local reader on two model families ($\Delta = +.132$, $p = .017$; $\Delta = +.198$, $p < .001$) while cutting false discoveries by 2–3 $\times$. Controlled replays attribute the gain and its limits: neither reference routing nor deterministic citation expansion alone exceeds the local baseline; whole-document reading (29–65K tokens), two-round iterative retrieval, and even oracle minimal-certificate evidence all fail to reach LASER, localizing the bottleneck in typed obligation closure rather than evidence access. An authority-corruption replay shows source-link errors only cause abstention, while the only unsafe-warrant mechanism is mis-recording a Held erratum as Verified, confined to that erratum’s family. LASER thus improves long-range dependency recovery while converting incomplete evidence and uncertain authority into explicit residual obligations rather than unsupported behavior.

1 Introduction

Internet standards are written for exposition and revision, not execution. Consider one real occurrence from RFC 4271 (BGP-4), used throughout this paper:

OPENCONFIRM state. “In response to any other event (*events 9, 12–13, 20, 27–28*), the local system: sends a NOTIFICATION message with the Error Code of Finite State Machine Error, . . . and changes its state to Idle.”

The clause names no event; it names six *numbers*. Events 9, 12, 13, 20, 27, and 28 are defined in a catalog 5,518 tokens away. The correct behavior is not one transition but six: $\text{OPENCONFIRM} \xrightarrow{\text{ConnectRetryTimer Expires?}} \text{IDLE}$,

*Corresponding author: hbzhu@sei.ecnu.edu.cn

DelayOpenTimer Expires?

→ IDLE, and so on, each with a mandatory sent NOTIFICATION. A reader that quotes only the local clause can guess “some event moves OpenConfirm to Idle”; it cannot justify *which* six events, nor bind each to its definition. Worse, a later erratum may replace this text entirely, so even a locally correct clause can be stale.

This paper treats such behavior recovery as *sparse defeasible evidence closure*. Evidence is *sparse* because the required passages occupy a small fraction of the document; *conjunctive* because every typed premise is required; and *defeasible* because a higher-priority correction can overturn local support. A longer context window does not establish closure: failing to retrieve a stronger rule is not evidence that the rule is absent.

Two kinds of obligation. Throughout the paper, $O^+(h)$ (*support obligations*) are the premises h still needs: a grounded source state, each triggering event, the target state, guards, updates. They are satisfied by *finding* evidence. $O^-(h)$ (*defeater obligations*) are the registered rules that could *overturn* h : errata, later versions, authority records. They are satisfied by *deciding* them—applicable or not, with authority established. The distinction is directional: absence of evidence leaves O^+ open; absence of a decision leaves O^- open, and an open O^- can never certify h . This is why LASER abstains *for a stated reason*, instead of merely abstaining with low confidence.

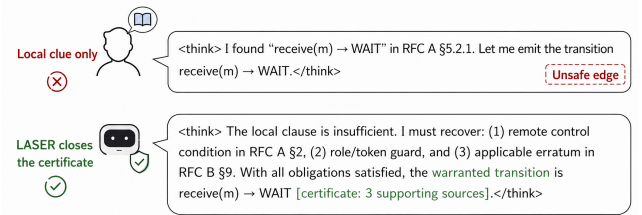


Figure 1: Local evidence versus certificate closure.

We first verify the problem in released artifacts: replaying RFCNLP annotations shows transition localization substantially overstates occurrence-aligned edge recovery (Section 3). LASER (*Long-range Acquisition of Sparse Evidence Routes*) then addresses the stricter task with a typed evidence

graph and proof-carrying search. A deterministic RuleIR checker alone returns **WARRANTED**, **REFUTED**, or **UNKNOWN**; learned or heuristic policies only order finite, graph-derived actions. Compilation to a timed-automata backend begins only after certificate closure.

Our contributions are:

- an artifact-based diagnosis showing that transition localization and occurrence-bound edge recovery diverge;
- a source-grounded, three-valued certificate semantics with an explicit treatment of the dual-closure case;
- an obligation-aware search algorithm with typed path actions, checker-exclusive verdicts, validated union, and certificate minimality, together with a fairness condition under which relative completeness holds;
- a family-paired evaluation with significance tests on two model families, a frozen-completion attribution replay separating routing from citation expansion, long-context, iterative-retrieval, and oracle-evidence controls, a protocol-level cue-transfer audit, and an authority-corruption replay that localizes the only unsafe-warrant mechanism.

2 Related Work

Protocol knowledge extraction. RFCNLP maps annotated protocol text to legacy FSMs and uses those models for attacker synthesis (Pacheco et al. 2022). PROSPER uses LLMs to extract protocol artifacts from document chunks (Sharma and Yegneswaran 2023); RFC2Code/APG turns an RFC-derived guidebook into code (Long, Su, and Han 2026); SecGoal targets formalizable security objectives rather than behavior machines (Huang et al. 2026). We reuse RFCNLP artifacts where a source-aligned diagnostic is possible. Their published scores are not placed in one leaderboard because phrase labels, code, security goals, and defeasible behavior certificates use different evaluation objects.

Proof-guided retrieval and verification. Closest to our acquisition interface, Aly and Vlachos retrieve documents autoregressively for fact verification and stop when a natural-logic proof over the retrieved set is sufficient (Aly and Vlachos 2022); ProofFVer executes symbolic natural-logic proofs to return verdicts (Krishna et al. 2022). LASER differs in the object being closed: occurrence-bound protocol premises with typed roles, cross-document bindings, and authority-sensitive defeaters whose status can be Verified, Held, or absent. There, evidence sufficiency is judged over a fixed claim and knowledge base; here, the claim itself is a behavior hypothesis whose premises are inherited from remote clauses and can be superseded by later corrections. This difference forces three-valued authority handling and residual-obligation output rather than proof-only verdicts.

Long-context and neuro-symbolic reasoning. Long-context models may underuse evidence even when it fits in the input (Liu et al. 2024; Bai et al. 2023). Retrieval-augmented generation supplies external evidence (Lewis et al. 2020), while multi-hop and adaptive systems interleave retrieval with reasoning (Xiong et al. 2021; Trivedi et al. 2023; Jeong et al. 2024; Sarthi et al. 2024), and agentic methods let a

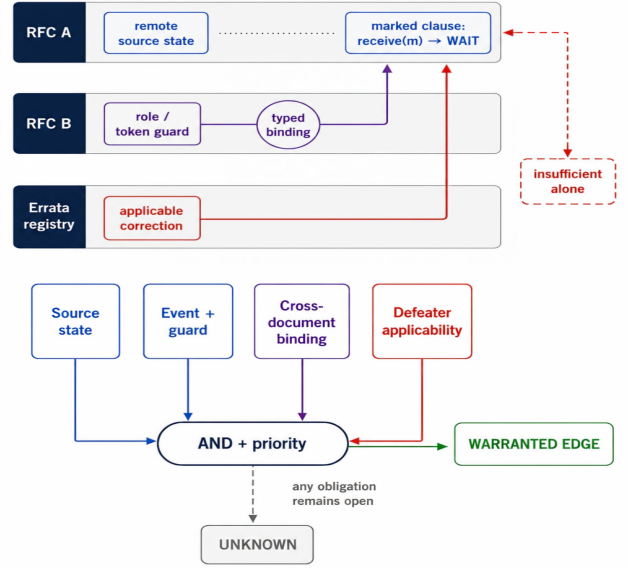


Figure 2: Sparse obligations behind a warranted edge.

model critique its own retrieval (Yao et al. 2023; Schick et al. 2023; Asai et al. 2024). These systems improve acquisition, but retrieved relevance alone does not establish that every support and defeater obligation has closed.

Natural-language reasoning benchmarks test rule application, proof production, and first-order inference (Clark, Tafjord, and Richardson 2020; Tafjord, Dalvi, and Clark 2021; Han et al. 2024); modular approaches expose intermediate steps or prover actions (Creswell, Shanahan, and Higgins 2023; Lyu et al. 2023; Chen et al. 2023; Yang et al. 2023), and Logic-LM and LINC delegate entailment to symbolic solvers (Pan et al. 2023; Olausson et al. 2023). LASER preserves the separation of authority but addresses an earlier, open-world problem: the premises must first be acquired across source routes, conjunctions, and possible defeaters. Its acceptance semantics instantiates finite defeasible reasoning (Reiter 1980; Maher 2001). Its **UNKNOWN** decision relates to selective prediction and semantic uncertainty (Geifman and El-Yaniv 2017; Kuhn, Gal, and Farquhar 2024; Wang et al. 2023), but follows explicit residual obligations rather than a confidence threshold.

3 Diagnosing Sparse Evidence Failures

3.1 From a transition clause to a certificate

A behavior hypothesis contains a source state, event, target state, participants, guard, update, clock, cross-document bindings, and an applicability audit. Its evidence is not a bag of relevant sentences: it is an **AND** certificate whose leaves must jointly witness the typed slots and resolve every registered stronger rule. Figure 2 makes this closure condition concrete; any open obligation forces **UNKNOWN** rather than an unsupported edge.

3.2 A released-artifact diagnosis

We test this distinction on RFCNLP’s released annotations and predictions. Predicted phrases are aligned to their source positions and passed through the released transition parser. For TCP, transition localization F1 is .804, but occurrence-aligned edge F1 is .427; for DCCP, .674 and .232. Required evidence occupies only .656% of RFC tokens at the median, yet the median certificate spans 384 tokens and the maximum 29,472. The inherited source state alone is 914 tokens away at the 90th percentile. Locating behavioral language does not recover the sparse control evidence needed to interpret it.

4 A Complete Certificate, End to End

Before defining the semantics, we walk through the OPENCONFIRM example on real data. Figure 3 shows the certificate LASER returns.

Occurrence. The marked occurrence is the clause “changes its state to Idle” inside the OPENCONFIRM block of RFC 4271 (anchor chunk at tokens 26,208–26,304). Its enclosing control block supplies the source state OPENCONFIRM and the mandatory sent message NOTIFICATION!.

Open support obligations. $O^+(h)$ contains one input-event obligation, but the occurrence names it by reference: “events 9, 12–13, 20, 27–28”. No evidence in the local window can close it. The reference route acquires the event catalog (tokens 19,584–20,736; up to 5,518 tokens away), where Event 9 is ConnectRetryTimer Expires, Event 12 is DelayOpenTimer Expires, Event 13 is IdleHoldTimer Expires, Event 20 is BGPOpen with DelayOpenTimer running, and Events 27–28 are UpdateMsg and UpdateMsgErr.

Defeater obligations. $O^-(h)$ contains the registered RFC Editor errata for RFC 4271. None binds to this occurrence, so the ledger Δ resolves each as inapplicable. Had a Verified Technical erratum replaced this clause, the same checker would return REFUTED for the stale behavior and WARRANTED only for the corrected one; a Held erratum would leave UNKNOWN.

Verdict and minimality. The typed expansion enumerates the six numbers into six atomic input events; the checker grounds each event label in its catalog entry, the source state in the enclosing control, the target in both the anchor chunk and the transition surface, and verifies the mandatory output ordering. All obligations close: the certificate contains 21 source chunks (2,016 tokens), and the output is six edges OPENCONFIRM→IDLE with their grounded inputs. A local reader on the same budget recovers none of the six.

5 LASER: Defeasible Certificate Search

Input and output. LASER receives a marked behavioral occurrence q , a public RFC bundle D , and a behavior hypothesis h . It returns a source-linked certificate for h , a certificate for an explicitly incompatible fragment, or UNKNOWN with the remaining obligations. The computation has three stages: normalization maps exact RFC spans, citations, section links, and errata into a typed evidence graph; a policy orders finite, graph-derived actions; and after every action

Certificate (abridged). occurrence: RFC4271 “changes its state to Idle” (t26208–26304) O^+: input-event[0] $\leftarrow$ reference “events 9, 12–13, 20, 27–28” Event 9 = ConnectRetryTimer Expires (t19584–19680) Event 12 = DelayOpenTimer Expires; Event 13 = IdleHoldTimer Expires (t19680–19776) Event 20 = BGPOpen w/ DelayOpenTimer (t20256–20352); Events 27,28 = UpdateMsg, UpdateMsgErr (t20640–20736) O^+: source $\leftarrow$ OPENCONFIRM control block; target $\leftarrow$ occurrence + surface; output $\leftarrow$ NOTIFICATION! O^-: RFC4271 errata $\rightarrow$ no binding to this occurrence Verdict: WARRANTED; 6 edges; 21 chunks, 2,016 tokens; leaf-minimal
--

Figure 3: The real certificate for the OPENCONFIRM example: marked occurrence, exact spans, resolved catalog bindings, applicability ledger, and verdict.

a symbolic checker recomputes support, applicability, and conflicts. The verdict depends only on selected source atoms and checker rules; learning affects search order only.

5.1 Source-grounded semantics

Let D contain immutable source atoms and their public relations. Each atom $e \in D$ records an exact document/section span and a typed interpretation (ℓ, τ, v, p) : literal content, semantic role, version, and priority. The normalizer constructs $G = (V, E_G)$, whose vertices are source atoms and document/section identifiers and whose typed edges encode adjacency, containment, citation, bibliography binding, version, and erratum relations. A behavior hypothesis h has slots for source state, event, target state, participants, guard, update, clock, cross-document binding, and applicability.

Definition 1 (Obligations) $O^+(h)$ is the finite set of typed support obligations of h . $O^-(h)$ is the set of higher-priority rules whose applicability must be established or ruled out. An evidence set E closes a support obligation only with a source-linked, type-compatible witness drawn from E ; absence from E never closes a defeater.

Let $\succ$ be an acyclic priority relation supplied by the source policy (e.g., an authoritative correction over its original text). A ledger Δ maps each $d \in O^-(h)$ to an applicable witness, an inapplicability witness, or *open*. Source-route and erratum bindings must be unique in the full source bundle, not merely within selected evidence. Thus a retrieved correction cannot defeat a clause unless the checker can bind both texts to the same occurrence and establish the correction’s authority.

Definition 2 (Defeasible certificate, evidence-relative)

Given evidence E , a certificate $P = (E_P, \Delta)$ for h is E -closed iff $E_P \subseteq E$, every $o \in O^+(h)$ has a consistent witness in E_P , every registered stronger rule is resolved by Δ , and all role, type, clock, version, and binding constraints pass. Write $\text{Closed}(h, E)$ iff some E -closed certificate for h exists.

The restriction $E_P \subseteq E$ is essential: the existential quantifier ranges only over selected evidence, so no conclusion can lean on unseen atoms. Writing $\text{Closed}(x, E)$ as $\text{Cl}(x)$ for

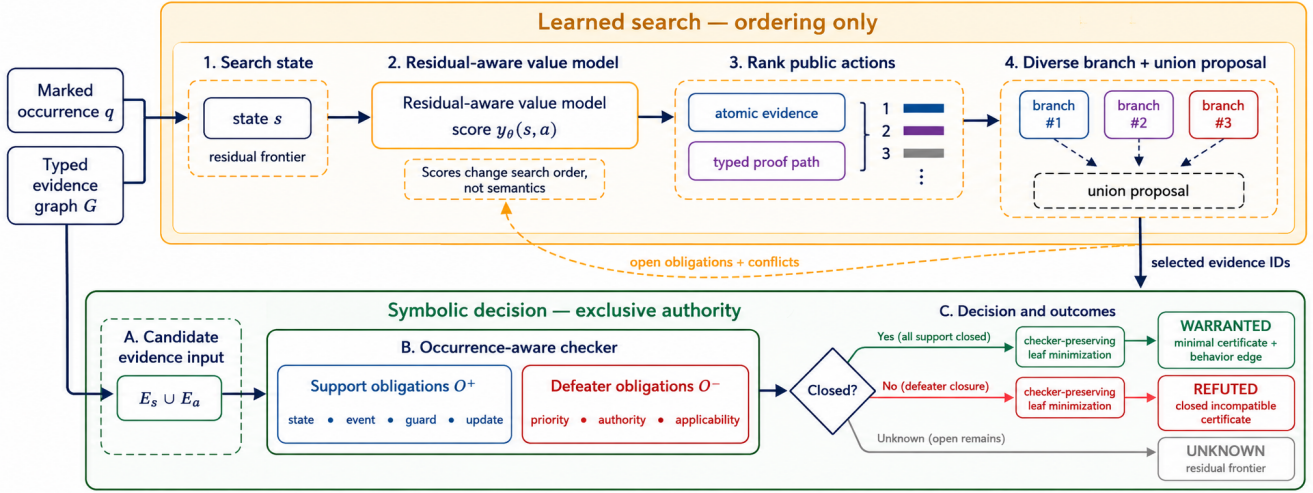


Figure 4: Learned search ordering with checker-exclusive warrant.

readability, the checker implements

$$\text{dec}(h, E) = \begin{cases} \text{WARRANTED} & \text{Cl}(h) \wedge \neg \text{Cl}(\bar{h}), \\ \text{REFUTED} & \text{Cl}(\bar{h}) \wedge \neg \text{Cl}(h), \\ \text{UNKNOWN}_{\text{conf}} & \text{Cl}(h) \wedge \text{Cl}(\bar{h}), \\ \text{UNKNOWN} & \text{otherwise,} \end{cases}$$

where $\bar{h}$ is an explicitly represented incompatible outcome, not an LLM-generated negation. Two clarifications follow. First, witnesses for h and $\bar{h}$ are drawn from the same E , so dual closure is a property of the evidence, not of search order. Second, $\text{UNKNOWN}_{\text{conflict}}$ is surfaced as UNKNOWN with an explicit conflict marker and both partial certificates; it is not silently resolved. Under acyclic $\succ$ and consistent authority records, a pair cannot both be closed by the same priority-resolved ledger; when authority records genuinely disagree, the conflict case is exactly the intended output.

The checker also returns feedback $z = (v, O_{\text{open}}^+, O_{\text{open}}^-, \Delta)$, where v is the four-valued internal decision collapsed to three output values. This feedback is the search interface: actions target named residual obligations, while the checker alone removes them. Progress is measured by closed obligations rather than sentence relevance. The guarantee is relative to the finite RuleIR; missing evidence can never be converted into support.

5.2 Proof-carrying evidence actions

Figure 4 shows the separation. The upper region orders finite public actions and forms union proposals from diverse branches. The lower region recomputes support and defeater closure, minimizes accepted certificates, and returns open obligations otherwise. An action $a = (E_a, k_a, u_a, c_a)$ contains exact evidence IDs E_a , an action kind k_a , a targeted obligation u_a , and token cost c_a . It contains no verdict. *Atomic actions* expose one source atom. *Path actions* expose one typed join, such as a source, bibliography binding, and target section. Paths can present mutually dependent leaves

in one expansion, but the checker still reconstructs every relation and rejects ambiguous or type-incompatible bindings. For selected evidence E_s , applying a means

$$s \oplus a = (E_s \cup E_a, \text{Check}(h, E_s \cup E_a)).$$

Consequently, an incorrect high-ranked action only consumes search budget; it cannot change the acceptance rule.

5.3 Obligation-aware branch-and-check

The action abstraction turns certificate completion into branch-and-check. A state $s = (E_s, z_s)$ stores selected evidence and the checker’s current feedback. At each expansion, LASER ranks only actions whose evidence is not already selected, creates individually rechecked children, and then creates a merge child from the evidence proposed for different open obligations. One child per residual-obligation type is retained before the remaining beam is filled. This diversity rule prevents several high-scoring support actions from crowding out the only action that investigates an applicable correction. States with the same evidence set are deduplicated; among unresolved states, fewer open obligations, lower evidence cost, and smaller depth define the deterministic queue order.

The evidence budget caps the unique selected-source tokens, the expansion budget caps checker calls to new search nodes, and the beam limits how many unresolved states survive each round. These are search resources, not confidence thresholds. Algorithm 1 is written for h ; $\bar{h}$ is checked under the same procedure. If neither side closes within budget, LASER returns the lowest-cost frontier with the fewest open obligations, which makes the reason for abstention inspectable.

5.4 Search values and conservative output

LASER supports three interchangeable action-ordering policies. *LASER-Symbolic* uses deterministic typed priority, *LASER-Distilled* fits a ridge value to checker-observed progress, and *LASER-LM* asks a fixed language model to rank

Algorithm 1 LASER branch-and-check

```

1:  $A \leftarrow \text{TypedActions}(G, h)$ 
2:  $z_0 \leftarrow \text{Check}(h, \emptyset); Q \leftarrow \{(\emptyset, z_0)\}$ 
3: while  $Q \neq \emptyset$  and budget remains do
4:    $s = (E_s, z_s) \leftarrow \text{PopBest}(Q)$ 
5:   if  $z_s.\text{verdict} = \text{WARRANTED}$  then
6:     return  $\text{Minimize}_{\text{Check}}(h, E_s)$ 
7:   end if
8:    $R \leftarrow \text{Rank}_{\theta}(\text{Available}(A, E_s) \mid h, z_s)$ 
9:    $B \leftarrow \text{DiverseTop}(R, z_s.O_{\text{open}}^+, z_s.O_{\text{open}}^-)$ 
10:  for  $a \in B$  do
11:    add  $s \oplus a$  to  $Q$ 
12:  end for
13:  add  $\text{CheckUnion}(s, B)$ ; deduplicate and prune
14: end while
15: return UNKNOWN with the best residual frontier

```

opaque public action IDs. Each policy receives the same action projections and residual obligations and returns only an ordering. The language model cannot invent a source atom, graph edge, or verdict.

The distilled policy uses process supervision rather than answer supervision. For a development state-action pair, its scalar target is

$$y_{\lambda}(s, a) = \lambda_{\text{close}} \mathbf{1}[z_{s \oplus a}.v = \text{WARRANTED}] + \Delta_O(s, a) - .02 \log(1 + c_a) - .01|E_a|,$$

where $\Delta_O(s, a) = |O_{\text{open}}(s)| - |O_{\text{open}}(s \oplus a)|$. The first term gives priority to a checker-closed child, the second rewards partial closure, and the last two break ties toward shorter evidence. The implementation fixes $\lambda_{\text{close}} = 8$. This dominance margin is scoped to the observed development statistics: across the 4,632 development actions used to fit the value head, the largest single-step obligation reduction is 3 and the largest cost penalty is below .197, so within those bounds 8 makes every checker-closed child outrank any merely partial child. Outside those bounds it remains a fixed lexicographic tie-break, and changing it cannot alter the checker’s verdict; it can only alter finite-budget action order.

The ridge features are lexical coverage, action kind, targeted residual type, source distance, and evidence cost. Targets use no expected answer or evaluation certificate. Before returning a proof, a fixed-point deletion pass tests every selected leaf and removes it whenever the same checker still returns WARRANTED. Thus the returned proof is inclusion-minimal, while learning changes only search efficiency and finite-budget recall.

5.5 Properties

Proposition 1 (Conservative soundness) *Relative to sound source links, types, and priority in a finite RuleIR, every LASER WARRANTED output has an E-closed certificate; an unresolved stronger rule cannot be treated as absent.*

Sketch. The checker is the sole constructor of a warrant. It quantifies over all registered in-scope higher-priority rules, requires a unique source binding, and restricts every witness

to the selected evidence E ; any failed predicate leaves a residual obligation.

Definition 3 (Retention fairness) *A search run is retention-fair iff, for every state s and action a available at s , either some child with evidence $E_s \cup E_a$ enters the queue at some round, or a state with evidence $E_s \cup E_a$ is discarded only in favor of a state with equal or fewer open obligations and no greater cost.*

Proposition 2 (Relative completeness under fairness)

With complete finite action enumeration, acyclic priority, unbounded budget, and retention fairness, LASER returns a closed certificate iff one exists in the candidate universe.

Sketch. Under retention fairness, no assignment containing a closing certificate is permanently discarded, so exhaustive expansion visits it; the deletion fixed point preserves warrant.

Scope of Proposition 2. The implemented beam and DiverseTop rules are *not* retention-fair in general: a finite beam can discard a state that a later closure would need. We therefore claim completeness only for the exhaustive, fair variant, and treat the implemented algorithm as an anytime procedure whose finite-budget misses are precisely the UNKNOWN residuals it reports. This is consistent with the conservative-output contract: a budget-limited UNKNOWN never asserts the absence of a certificate.

Proposition 3 (Minimality) *Every returned certificate is leaf-minimal under the checker: removing any remaining leaf destroys warrant.*

Sketch. The deletion pass reaches a fixed point; by construction every surviving leaf was tested and found necessary. This is inclusion minimality, not globally minimum token cost.

6 Experiments

6.1 Datasets and Tasks

BehaviorCheck. We derive 103 occurrence-bound behavior families from released BGPv4, DCCP, SCTP, and TCP annotations. Twelve families support task and prompt development; the remaining 91 form the evaluation split. SCTP contributes 13 source transitions, all of which the occurrence-aware projection excludes (10 lack an inheritable source state, three lack a required slot), so the evaluation split contains BGPv4 (52), TCP (31), and DCCP (8) families; SCTP remains in the projection audit. Each family requires an atomic source, event, and target edge together with the evidence that warrants it. Twenty-eight families require a numbered or complementary reference outside the local transition context.

Defeater and authority evaluation. DefeaterCheck contains 31 real RFC Editor erratum routes. Twenty-five have one applicable Verified Technical correction; six remain unresolved. A separate authority evaluation contains 20 paired stale and corrected claims (40 decisions) from six protocol families: 12 Verified and 8 Held. We also use 277 controlled long-range routes (110 positive, 167 negative) to measure search behavior under fixed distractors.

Reader	Edge recovery					Proof quality			Decision	
	P	R	F_1	Macro F_1	Exact	Evid. F_1	nDCG	Joint	FDR ↓	Abstain
<i>Local context</i>										
Previous context	.647	.219	.327	.302	.297	.406	.412	.220	.353	.538
BM25 + source neighbors	.581	.285	.382	.313	.308	.486	.474	.242	.419	.462
Occurrence-centered context	.582	.258	.358	.374	.363	.559	.554	.308	.418	.407
<i>Long-range evidence routing</i>										
LASER (reference)	.821	.609	.700	.618	.593	.684	.623	.407	.179	.242
LASER (hybrid)	.813	.576	.674	.630	.626	.692	.646	.440	.187	.242
LASER (cue-gated)	.810	.623	.704	.604	.604	.668	.584	.418	.190	.286

Table 1: Behavior recovery under a matched 2K evidence budget.

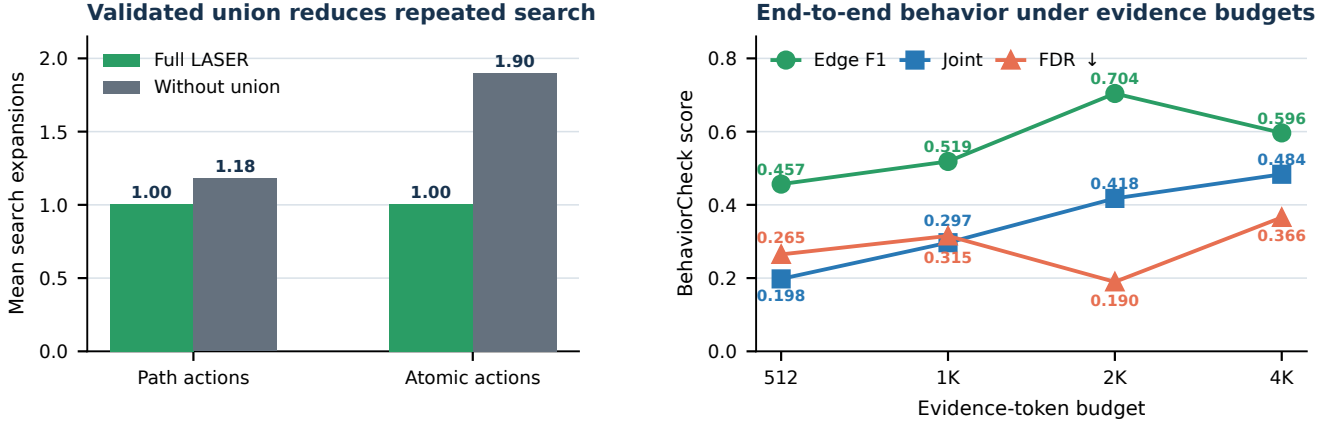


Figure 5: Component ablation and evidence-budget response.

Executable backend. Closed CoAP timeout certificates are compiled to a timed-automata interface and checked against implementation traces and UPPAAL queries. This case study evaluates the compilation boundary, not natural-language evidence labels.

6.2 Experimental Settings

All behavior readers use the same language model, candidate anchor, structured output contract, occurrence-aware checker, and 2K evidence budget. The strongest local baseline on Joint Edge+Proof is the occurrence-centered reader; the strongest on micro Edge F1 is BM25 + source neighbors; both appear in Table 1. LASER adds two fixed routing policies: a cue-gated policy uses reference search only when a public dependency cue is present, while a hybrid policy combines local and reference evidence. On errata, atomic and path actions share the same checker. We report micro and macro Edge F1, exact edge-set match, evidence F1 and nDCG, Joint Edge+Proof, false discovery rate (complement of precision, reported for continuity), abstention, three-valued accuracy, and unsafe warrants. Prompts, search limits, and remaining hyperparameters are in the supplementary material; every run records prompt and completion hashes.

6.3 Main Results

Table 1 reports the matched behavior comparison. Against the strongest local Joint reader (occurrence-centered, .308),

the LASER hybrid reaches .440 Joint Edge+Proof; against the strongest local Edge-F1 reader (BM25 + source neighbors, .382), the cue-gated policy reaches .704. Family-paired statistics appear in Table 5: the hybrid improves Joint by +.132 (95% CI [+ .033, +.231], permutation $p = .017$; Wilcoxon $p = .011$), exact match by +.264 and Edge F1 by +.256 (both $p < 10^{-3}$). The cue-gated and reference policies improve Edge F1 and exact match significantly ($p < .005$) with positive but marginal Joint deltas ($p = .096$ and $.121$), reflecting the smaller hybrid edge-set improvements on local-route families.

Table 3 localizes the gain. On 28 reference-required families, Edge F1 rises from .025 to .726, exact match from .000 to .643, and Joint from .000 to .321. Throughout, “Local” denotes the BM25 + source-neighbors reader and “LASER” the cue-gated configuration of Table 1. BGPv4 improves from .188 to .780 Edge F1; DCCP and TCP are unchanged because the cue gate retains local reading when no remote route is exposed.

6.4 What Produces the Gain? An Attribution Replay

A reviewer can reasonably ask whether the remote-reference gains come from obligation-aware routing or merely from a deterministic citation-expansion rule. We therefore replay *frozen completions* through the same checker in a 2×2 design (Table 4): reference routing on/off $\times$ deterministic expansion

Removed component	Evaluation	Δ correct/closed	Δ expansions	Safety effect
Proof paths	25 applicable errata	−25 Closed@1	+0.87	No unsafe change
Checker feedback	277 controlled routes	0	0.00	0/167 false routes
Diverse beam	277 controlled routes	0	0.00	0/167 false routes
Validated union	277 controlled routes	0	+0.18 path; +0.90 atomic	0/167 false routes
Authority checking	20 paired families	−16 correct claims	N/A	+8 unsafe warrants

Table 2: Removal ablation across controlled evaluations.

Slice	n	Edge F_1			Exact match		Joint Edge+Proof	
		Local	LASER	Δ	Local	LASER	Local	LASER
<i>Dependency regime</i>								
Remote reference	28	.025	.726	+.701	.000	.643	.000	.321
Local route	63	.579	.688	+.109	.444	.587	.349	.460
<i>Protocol</i>								
BGPv4	52	.188	.780	+.592	.173	.692	.096	.404
DCCP	8	.769	.769	.000	.625	.625	.625	.625
TCP	31	.568	.568	.000	.452	.452	.387	.387

Table 3: Behavior recovery by dependency regime and protocol.

Routing	Expansion	Edge F_1	Joint	Remote F_1
Local	off	.382	.242	.025
Local	on	.382	.242	.025
Reference	off	.396	.253	.054
Reference	on	.700	.407	.719

Table 4: Attributing the remote-reference gains by replaying frozen completions through the same checker. Neither deterministic citation expansion without routing nor routing without the typed expansion closes remote references; the gain requires both components. Remote column: Edge F_1 on the 28 reference-required families.

on/off. Applying the expansion rule to the local reader’s evidence changes nothing (Edge F_1 stays .382, remote .025): the rule cannot fire because the event catalog is not in the budget. Conversely, routing without the typed expansion reaches only .396 overall and .054 remote: the model sees the catalog but does not reliably enumerate the referenced numbers. Only the combination closes remote references (.700 overall, .719 remote). Neither component alone differs significantly from the local baseline (Table 5). The gain is therefore attributable to the interaction—routing acquires the typed catalog evidence, and the checker-side expansion closes the numeric obligation—not to either mechanism in isolation.

6.5 Cue-Gate Transfer Across Protocols

Table 6 audits the public cue gate per protocol. The gate is a fixed, prediction-blind pattern test (numbered event reference or explicit complement); gold labels are used only to evaluate it. It fires on all 52 BGPv4 families (recall 1.0), on one of two required DCCP families, and never on TCP, missing one required TCP complement variant. Overall recall is .93 with precision .49: the gate over-fires on BGPv4, where reference

Method	Δ Joint	CI95	p	Δ Edge
Hybrid	+0.132*	[0.033,0.231]	0.017	+0.256
Cue-gated	+0.110	[-0.011,0.220]	0.096	+0.230
Reference	+0.099	[-0.011,0.209]	0.121	+0.244
Expansion only	-0.066	[-0.143,0.011]	0.182	-0.061
Routing only	-0.055	[-0.132,0.022]	0.262	-0.023

Table 5: Family-paired LASER variants and attribution arms against the strongest local reader (occurrence-centered) on Joint Edge+Proof, 20,000 bootstrap and permutation rounds, family as the unit ($n=91$); p : two-sided permutation test, * $p < 0.05$.

routing never removes a correct local edge, and its misses are exactly two uncovered complement phrasings. This delimits where the current routing policy transfers: pattern-level generalization, with recall bounded by public cue coverage.

6.6 Second-Model Replication, Long-Context, Iterative, and Oracle Controls

Table 7 reports a full second-model replication (GLM-5.3, all 91 families, same prompt, checker, and 2K budget) with four additional controls. First, the routing-over-local ordering replicates with larger margins than on the primary model: LASER hybrid reaches .803 Edge F_1 / .473 Joint versus .447 / .275 for the strongest local reader (family-paired Δ Joint = +.198, $p < .001$; only 2 of 91 families regress). Second, *whole-document reading is the weakest reader of all* (.276 Edge F_1 , .545 false discovery rate) despite 29K–65K tokens of evidence: unrestricted context adds distractors without telling the model *which* obligations to close. Third, a two-round checker-residual-driven iterative reader under the same total budget does not help (.264), showing that generic adaptive retrieval is not the missing ingredient. Fourth, an

Protocol	Fam.	Fired	Req.	Prec.	Rec.
BGPv4	52	52	25	.48	1.00
DCCP	8	1	2	1.00	.50
TCP	31	0	1	.00	.00
All	91	53	28	.49	.93

Table 6: Public cue-gate transfer audit; Req.: reference required, Prec./Rec.: cue precision/recall. The gate is a fixed, prediction-blind pattern test; gold is used only to evaluate it. It misses two required references (one DCCP and one TCP complement variant) and over-fires on BGPv4, where reference routing never removes a correct local edge.

oracle arm that selects the gold minimal-certificate evidence (labels never shown) reaches only .496: the minimal certificate does not include the catalog chunks that typed expansion needs, so even perfect retrieval of the recorded certificate cannot match routing that acquires the catalog. Together these controls localize the remaining bottleneck precisely where LASER acts—typed obligation closure—rather than in evidence access, budget, or model scale.

6.7 Sensitivity to Authority and Source-Link Errors

Proposition 1 is relative to correct links, types, and priorities, so we measure what happens when that assumption fails. Table 8 replays the deterministic laser-path method on the 20-family authority suite while corrupting one erratum record at a time. The clean replay reproduces all 40 expected labels. Three findings follow. (1) Corrupting source links (RFC binding or stale text) *never* creates an unsafe warrant; it only converts closures into UNKNOWN, so the failure mode is conservative abstention. (2) The only corruption that creates unsafe warrants is recording a Held erratum as Verified; all 16 unsafe warrants come from the 8 Held families, each corruption affecting exactly its own family. (3) Flipping Verified to Held only loses closures (24), again without unsafe warrants. End-to-end soundness therefore decomposes: representation errors degrade to abstention, and authority-status errors are the single, localized unsafe-warrant channel.

6.8 Ablation Study

Table 2 reports matched removal effects. Removing proof paths loses all 25 one-expansion closures; removing validated union increases search cost, especially for atomic actions. Checker feedback and beam diversity show no measurable effect on the controlled routes; this is expected rather than anomalous. On those routes the symbolic priority is already near-ceiling: on the 277-route controlled suite, the symbolic policy alone reaches .991 positive certificate recall with zero model calls and a mean of 55 evidence tokens, leaving learned ranking nothing to improve. The situation differs on BehaviorCheck, where the routing policy itself—which evidence family to consult—carries the gain (Section 6.4). The practical benefit of learning is thus scoped: it can only reorder finite actions, so it helps exactly where the symbolic order is suboptimal, and our current suites bound that

regime. Authority checking is essential for safety: removing it loses 16 correct claim decisions and turns all eight Held corrections into unsafe warrants.

Figure 5 connects the search ablation to the end-to-end budget response. Moving from 512 tokens to 2K recovers missing support; at 4K, additional proof leaves improve Joint Edge+Proof while distractor edges degrade Edge F1 and precision. LASER benefits from targeted acquisition rather than an unrestricted context window.

6.9 RuleIR Construction Audit

The BehaviorCheck projection audit covers 162 source transitions: 103 are included and 59 are excluded with machine-checkable reasons (25 ambiguous or missing target, 18 missing source state, 9 missing input event, 3 no control context, 4 requiring review). On the sealed authority suite, all 20 erratum stale texts have a unique occurrence in their frozen RFC, so the binding-uniqueness precondition holds. In the reference run, the typed resolver finds a complete public catalog for 47 of 49 numbered-reference families (the 2 incomplete cases abstain), plus 1 complement family. These audits quantify how the finite RuleIR is built and where it abstains; they do not replace independent annotation review, which remains a stated limitation.

Checker robustness. Table 9 replays frozen GLM completions through the occurrence checker while sweeping its window constants and removing its alias table. Results saturate at $\pm 192/96$ tokens and are unchanged by the aliases or a doubled window; even a four-fold tighter window keeps precision at .880 while losing recall. The checker’s constants are therefore not hidden tuning; below saturation it degrades conservatively, and above saturation they are inert.

6.10 Analysis

The gains follow dependency structure. The improvement concentrates on reference-required families and the protocol ordering follows measured control distance, supporting the claim that LASER repairs sparse remote dependencies rather than uniformly changing predictions.

Search and decision authority remain separate. Paths and validated union reduce the expansions needed to assemble a conjunction; all compared policies share the same occurrence, type, priority, and authority checks, which explains why finite-budget closure improves without increasing unsafe warrants.

Executable interface. The compiled CoAP model passes two pinned traces, two multicast queries, and four terminal-aware UPPAAL queries. Query results come from verifier logs; the case study validates the compilation boundary, not natural-language interpretation.

6.11 Limitations and Validity

Primary-model results use frozen completions re-used across replays; the full 91-family GLM-5.3 replication (Table 7) covers the second model family, but two families of models do not establish model-independent calibration. The suites

Reader (GLM-5.3, $n=91$)	Evidence	Edge F_1	Joint	Exact	Δ Joint	p
Iterative retrieval (2 rounds)	2K	.264	.176	.242	-0.099	0.065
Direct-full (whole document)	29–65K	.276	.209	.253	-0.066	0.240
Local BM25 + source neighbors	2K	.406	.275	.319	+0.000	1.000
Occurrence-centered (strongest local)	2K	.447	.275	.352	—	—
Oracle certificate evidence	$\leq 2K$	.496	.330	.407	+0.055	0.474
LASER (reference)	2K	.799	.484	.626	+0.209	<.001
LASER (cue-gated)	2K	.791	.473	.626	+0.198	<.001
LASER (hybrid)	2K	.803	.473	.626	+0.198	<.001

Table 7: Full second-model replication on GLM-5.3 (all 91 families; same prompt, checker, and 2K budget). Δ Joint and p are family-paired against the strongest local reader. Whole-document reading, two-round iterative retrieval, and even oracle minimal-certificate evidence do not reach LASER: the bottleneck is typed obligation closure, not evidence access. Oracle uses private certificate evidence for selection only.

Corruption	Records	Unsafe	Lost
Held $\rightarrow$ Verified	8	16	0
Verified $\rightarrow$ Held	12	0	24
RFC binding corrupted	20	0	24
Stale text corrupted	20	0	24

Table 8: Joint authority/source-link corruption replay on the 20-family sealed authority suite (40 claims). Clean replay reproduces all 40 expected labels. Unsafe: unsafe warrants; Lost: lost closures. Source-link and binding corruption never creates an unsafe warrant; it degrades to UNKNOWN. Mis-recording a Held erratum as Verified is the only unsafe-warrant mechanism, and it affects exactly the corrupted family.

Window	Warranted	Edges	Edge F_1	Joint
$\pm 64/32$	59	92	.667	.418
$\pm 128/64$	67	121	—	—
$\pm 192/96$	68	122	.799	.484
$\pm 384/192$	None	None	.799	.484
$\pm 768/384$	68	122	—	—
no aliases	68	122	.799	.484

Table 9: Occurrence-checker sensitivity replay over frozen GLM completions; the deployed window is $\pm 384/192$. Results saturate at $\pm 192/96$ and are unchanged by the alias table or a doubled window; halving the window to $\pm 64/32$ loses recall while keeping precision (.880), i.e., degradation is conservative.

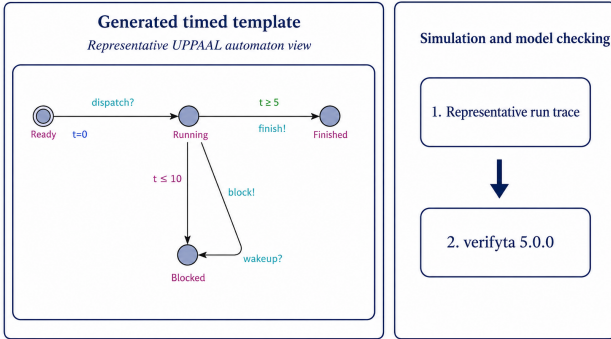


Figure 6: UPPAAL model and verification workflow.

are development-grade: family labels come from a deterministic projection pending independent annotation review. External validity is limited to three evaluated protocols and 91 families; the cue gate generalizes at the pattern level and its recall is bounded by public cue coverage (Table 6). Construct validity inherits the released annotation schema; checker soundness is relative to the finite RuleIR, and Section 6.7 bounds the damage of its two error classes. The oracle arm reveals a benchmark artifact of its own: recorded minimal certificates omit catalog evidence, so the oracle is a lower-style ceiling on certificate-only retrieval rather than a

true upper bound. FDR is the complement of precision and is retained only for continuity with prior reporting; unsafe-warrant behavior is measured on the authority suite. The UPPAAL case study validates compilation, not extraction. Broader protocols, independent annotation, and calibration remain future work.

7 Conclusion

LASER reframes protocol behavior recovery as defeasible certificate search rather than local clause extraction. Its central contribution is the interface between learned acquisition and symbolic warrant: typed routes gather sparse, mutually dependent evidence; policies decide where to search; and an occurrence-aware checker retains exclusive authority over support, conflict, and correction applicability. The revision strengthens this claim with three audit mechanisms: family-paired significance tests, a frozen-completion attribution replay showing the remote-reference gains require routing and typed expansion jointly, and an authority-corruption replay showing that source-link errors degrade to abstention while the single unsafe-warrant channel is localized to authority-status records. Unresolved errata and uncertain authority remain explicit rather than becoming unsupported transitions.

Evaluation	Metric	Comparator	LASER
Applicable errata	Closed@1	0/25	25/25
	Mean expansions	1.87	1.00
Unresolved errata	Unsafe warrant	0/6	0/6
Authority claims	Correct claims	38/40	40/40
	Correct pairs	18/20	20/20
Held authority	Unsafe warrant	0/8	0/8

Table 10: Outcomes beyond BehaviorCheck.

References

- Aly, R.; and Vlachos, A. 2022. Natural Logic-guided Autoregressive Multi-hop Document Retrieval for Fact Verification. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 6284–6297.
- Asai, A.; Wu, Z.; Wang, Y.; Sil, A.; and Hajishirzi, H. 2024. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. In *International Conference on Learning Representations*.
- Bai, Y.; Lv, X.; Zhang, J.; Lyu, H.; Tang, J.; Huang, Z.; Du, Z.; Liu, X.; Zeng, A.; Hou, L.; Dong, Y.; and Li, J. 2023. LongBench: A Bilingual, Multitask Benchmark for Long Context Understanding. *arXiv preprint arXiv:2308.14508*.
- Chen, W.; Ma, X.; Wang, X.; and Cohen, W. W. 2023. Program of Thoughts Prompting: Disentangling Computation from Reasoning for Numerical Reasoning Tasks. *Transactions on Machine Learning Research*.
- Clark, P.; Tafjord, O.; and Richardson, K. 2020. Transformers as Soft Reasoners over Language. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, 3882–3890.
- Creswell, A.; Shanahan, M.; and Higgins, I. 2023. Selection-Inference: Exploiting Large Language Models for Interpretable Logical Reasoning. In *International Conference on Learning Representations*.
- Geifman, Y.; and El-Yaniv, R. 2017. Selective Classification for Deep Neural Networks. In *Advances in Neural Information Processing Systems*, volume 30, 4878–4887.
- Han, S.; Schoelkopf, H.; Zhao, Y.; Qi, Z.; Riddell, M.; et al. 2024. FOLIO: Natural Language Reasoning with First-Order Logic. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 22017–22031.
- Huang, D.; Li, H.; Jia, B.; Feng, H.; Guan, J.; Jiao, Y.; and Li, X. 2026. SecGoal: A Benchmark for Extracting Formalizable Security Goals from Protocol Documents. *arXiv preprint arXiv:2604.27601*.
- Jeong, S.; Baek, J.; Cho, S.; Hwang, S. J.; and Park, J. C. 2024. Adaptive-RAG: Learning to Adapt Retrieval-Augmented Large Language Models through Question Complexity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.
- Krishna, V.; Wu, X.; Riedel, S.; and Stenetorp, P. 2022. ProofVer: Natural Logic Theorem Proving for Fact Verification. In *Findings of the Association for Computational Linguistics: ACL 2022*, 2559–2573.
- Kuhn, L.; Gal, Y.; and Farquhar, S. 2024. Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation. *Nature*, 630: 625–630.
- Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Kuttler, H.; Lewis, M.; Yih, W.-t.; Rocktaschel, T.; Riedel, S.; and Kiela, D. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, volume 33, 9459–9474.
- Liu, N. F.; Lin, K.; Hewitt, J.; Paranjape, A.; Bevilacqua, M.; Petroni, F.; and Liang, P. 2024. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics*, 12: 157–173.
- Long, J.; Su, J.; and Han, B. 2026. LLMs Unleashed: Generating Protocol Code from RFC Specifications. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, 809–817.
- Lyu, Q.; Havaldar, S.; Stein, A.; Zhang, L.; Rao, D.; Wong, E.; Apidianaki, M.; and Callison-Burch, C. 2023. Faithful Chain-of-Thought Reasoning. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*.
- Maier, M. J. 2001. Propositional Defeasible Logic Has Linear Complexity. *Theory and Practice of Logic Programming*, 1(6): 691–711.
- Olausson, T. X.; Gu, A.; Lipkin, B.; Zhang, C. E.; Solar-Lezama, A.; Tenenbaum, J. B.; and Levy, R. 2023. LINC: A Neurosymbolic Approach for Logical Reasoning by Combining Language Models with First-Order Logic Provers. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 5153–5176.
- Pacheco, M. L.; von Hippel, M.; Weintraub, B.; Goldwasser, D.; and Nita-Rotaru, C. 2022. Automated Attack Synthesis by Extracting Finite State Machines from Protocol Specification Documents. In *2022 IEEE Symposium on Security and Privacy*, 51–68.
- Pan, L.; Albalak, A.; Wang, X.; and Wang, W. Y. 2023. Logic-LM: Empowering Large Language Models with Symbolic Solvers for Faithful Logical Reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 3806–3824.
- Reiter, R. 1980. A Logic for Default Reasoning. *Artificial Intelligence*, 13(1–2): 81–132.
- Sarathi, P.; Abdullah, S.; Tuli, A.; Khanna, S.; Goldie, A.; and Manning, C. D. 2024. RAPTOR: Recursive Abstractive Processing for Tree-Organized Retrieval. In *International Conference on Learning Representations*.
- Schick, T.; Dwivedi-Yu, J.; Dessì, R.; Raileanu, R.; Lomeli, M.; Hambro, E.; Zettlemoyer, L.; Cancedda, N.; and Scialom, T. 2023. Toolformer: Language Models Can Teach Themselves to Use Tools. In *Advances in Neural Information Processing Systems*, volume 36, 68539–68551.
- Sharma, P.; and Yegneswaran, V. 2023. PROSPER. In *Proceedings of the 22nd ACM Workshop on Hot Topics in Networks*, 41–47. ACM.
- Tafjord, O.; Dalvi, B.; and Clark, P. 2021. ProofWriter: Generating Implications, Proofs, and Abductive Statements over Natural Language. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 3621–3634.
- Trivedi, H.; Balasubramanian, N.; Khot, T.; and Sabharwal, A. 2023. Interleaving Retrieval with Chain-of-Thought Reasoning for Knowledge-Intensive Multi-Step Questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 10014–10037.
- Wang, X.; Wei, J.; Schuurmans, D.; Le, Q. V.; Chi, E. H.; Narang, S.; Chowdhery, A.; and Zhou, D. 2023. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *International Conference on Learning Representations*.

Xiong, W.; Li, X. L.; Iyer, S.; Du, J.; Lewis, P.; Wang, W. Y.; Mehdad, Y.; Yih, W.-t.; Riedel, S.; Kiela, D.; and Oguz, B. 2021. Answering Complex Open-Domain Questions with Multi-Hop Dense Retrieval. In *International Conference on Learning Representations*.

Yang, K.; Swope, A.; Gu, A.; Chalamala, R.; Song, P.; Yu, S.; Godil, S.; Prenger, R. J.; and Anandkumar, A. 2023. LeanDojo: Theorem Proving with Retrieval-Augmented Language Models. In *Advances in Neural Information Processing Systems*, volume 36, 21573–21612.

Yao, S.; Zhao, J.; Yu, D.; Du, N.; Shafran, I.; Narasimhan, K. R.; and Cao, Y. 2023. ReAct: Synergizing Reasoning and Acting in Language Models. In *International Conference on Learning Representations*.