WHEN EVIDENCE CHANGES THE SUBJECT: SUBJECT-TYPED CLAIM LICENSING FOR LEARNED ROUTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern learned systems increasingly combine learned components with search, repair, or external solvers. Benchmarks often measure the resulting end-to-end system, while scientific claims may concern only one component, creating an attribution problem: evidence can fail to support the requested component-level claim while still supporting a positive conclusion about the larger system. Existing evidence-to-claim methods primarily calibrate claim strength. We argue that composite systems require a second dimension: *scientific subject*. We address this problem with *subject-typed claim licensing*, which separates weaker conclusions about the requested subject from positive but non-substitutive credit about another subject. We instantiate this idea in SCOPE-ROUTING for preference-conditioned multigraph routing. Non-authors reproducibly apply the declared semantics; held-out review yields fewer reference-relative upward deviations than unstructured review, while the difference from a strong evidence checklist remains unresolved; and a controlled routing study shows that score-optimal and claim-eligible methods can differ while valid hybrid-system credit is preserved. These results motivate treating claim strength and scientific subject as distinct dimensions of evidence-based evaluation.

1 INTRODUCTION

Modern learned systems are increasingly composite: a learned component may operate with search or repair (Li et al., 2025), external tools or retrieval (Qin et al., 2024; Asai et al., 2024), or downstream experts (Ghoummaid & Shalit, 2024; Mao et al., 2024), while benchmark evaluation measures the behavior of the complete executed system. This creates a scientific attribution problem. As Figure 1 illustrates, two routers can achieve similar end-to-end regret even when one solves requests through its learned policy and the other relies heavily on a backup solver. The scores may be similar, but the evidence supports claims about different subjects.

This distinction creates two forms of *partial support*. Evidence may support a weaker conclusion about the *same scientific subject*—for example, static rather than dynamically robust learned routing (Xiao et al., 2024; Heakl et al., 2026). Or it may support a positive conclusion about a *different subject*—for example, the backup-enabled system rather than the learned policy. The first changes claim strength; the second changes what the supported claim is about. Collapsing the two can either misattribute system-level evidence to a component or discard genuine positive evidence about the system.

Recent work has made evidence-to-claim reasoning increasingly explicit. Benchmark-assessment and metadata standards expose evaluation conditions, lifecycle, and scope (Reuel et al., 2024; Akhtar et al., 2024); recent claim-verification methods assess whether designated statements are grounded in evidence, including claims drawn from scientific literature (Tang et al., 2024; Jansen et al., 2025); and evidence-licensed calibration and claim replay formalize how evidence constrains the strength of reportable claims (Li, 2026). These approaches clarify *how strongly* a claim is supported. Composite systems expose an additional representational question: *which scientific subject is supported?* When positive evidence concerns another component or the executed system, it need not lie on the weaker-to-stronger scale of the originally requested component claim.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

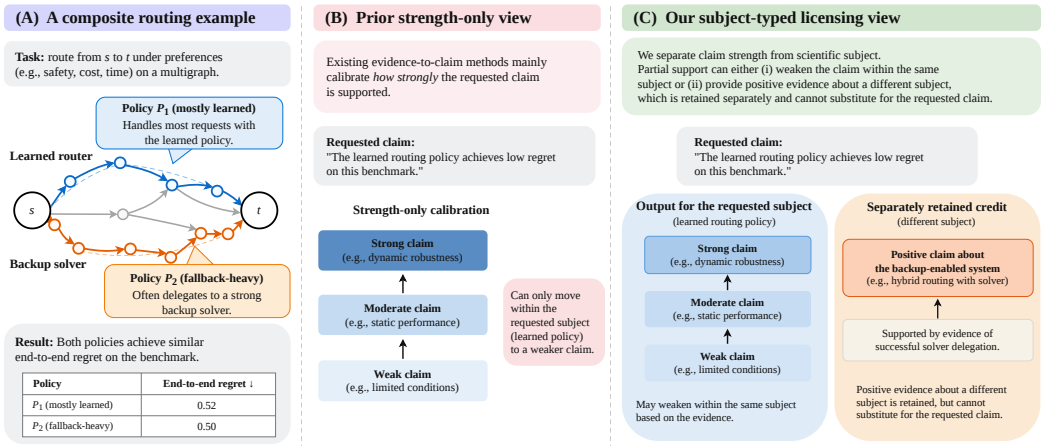

Figure 1: When evidence changes the subject. A fallback-heavy router illustrates the difference between strength-only calibration and subject-typed licensing. Strength-only calibration can only move along the requested-subject ladder; subject-typed licensing additionally retains positive evidence about the backup-enabled hybrid without allowing that evidence to support the standalone learned-router claim.

We address this problem with *subject-typed claim licensing*. The central idea is to separate claim strength from scientific subject. A requested claim may weaken only within the same subject, while positive evidence about a related but incomparable subject is retained separately and cannot substitute for evidence about the requested subject. We instantiate this semantics in SCOPE-ROUTING for preference-conditioned multigraph routing. Routing is a useful testbed because repair, multi-expert solver composition, edge identity, preference coverage, and dynamic response make attribution and claim scope directly observable (Li et al., 2025; Pan et al., 2026; Rydin et al., 2026; Chen et al., 2025; Xiao et al., 2024; Heakl et al., 2026).

We evaluate the approach along the path from scientific text to scientific decision. First, we test whether non-authors can reconstruct the required claim–evidence objects and apply a frozen contract consistently. Second, we compare held-out expert judgments under careful review, a strong evidence checklist, and subject-typed licensing. Third, we compare score-only method selection with selection conditioned on a requested scientific claim. Non-authors reproducibly apply the declared semantics; subject-typed review yields fewer reference-relative upward deviations than unstructured review, while its difference from the checklist remains unresolved; and the controlled routing study shows that score minimization and target-claim eligibility can select different policies while preserving valid hybrid-system credit.

Contributions. Conceptually, we identify scientific subject as a distinct dimension of evidence-relative claim interpretation. Formally, we introduce a subject-typed licensing semantics that separates same-subject narrowing from non-substitutive cross-subject credit. Empirically, we study its reconstruction, execution, reviewer consequences, and claim-conditioned selection in learned routing.

2 RELATED WORK

Evidence records, validity, and claim support. Evaluation methodology first determines what evidence is available and what interpretations that evidence can sustain. Decision-quality metrics and stress tests characterize observed behavior, while datasheets, model cards, BenchmarkCards, behavioral evaluation, benchmark-assessment frameworks, and reproducibility standards document conditions, intended uses, artifacts, and failure modes (Gebru et al., 2021; Mitchell et al., 2019; Ribeiro et al., 2020; Pineau et al., 2021; Sokol et al., 2026; Reuel et al., 2024). Validity-centered approaches ask which interpretations of observed measurements are warranted, and structured assurance approaches make claims, assumptions, evidence, and support relations explicit (Kane, 2013; Rhodes et al., 2010; Freiesleben & Zezulka, 2025; Salaudeen et al., 2025; Bean et al., 2026).

Recent claim-assessment systems additionally retrieve evidence, evaluate claim–evidence relations, or estimate support and overstatement for designated scientific statements (Javaji et al., 2025; James et al., 2026; Yue et al., 2026).

Evidence-licensed calibration and claim replay. Evidence-licensed calibration studies which assertion strength is justified by an evidence record, while claim replay makes evidence-relative inference executable over frozen evaluation records (Li, 2026). We take this claim-relative perspective as our starting point. Our focus is the output semantics induced by composite systems. When evidence about a requested component also licenses a conclusion about another component or the executed system, those conclusions have different scientific roles: one answers the requested question, while the other is valid positive credit about a different subject. Prior claim-relative frameworks can in principle be extended with additional typed conclusions; subject-typed licensing makes this distinction explicit through subject-preserving claim strength, a requested-subject output, relation-labeled retained credit, and a non-substitution rule.

Learned routing as an attribution testbed. Neural combinatorial optimization and decision-focused learning evaluate downstream decision quality, feasibility, robustness, and efficiency (Bello et al., 2016; Kool et al., 2018; Elmachtoub & Grigas, 2022; Mandi et al., 2024; Zhou et al., 2024). Multiobjective and multigraph routing additionally make preference coverage, action identity, and structural fidelity part of the evaluated object (Deb et al., 2002; Chen et al., 2025; Rydin et al., 2026). Repair, solver delegation, and stochastic requests create cases in which similar aggregate scores arise from different component contributions (Li et al., 2021; 2025; Heakl et al., 2026). Routing therefore provides request-level observations of attribution, task identity, coverage, dynamics, and intervention. Appendix A gives feature-level comparisons with adjacent evaluation and claim-support frameworks.

3 Subject-Typed Claim Licensing

We now formalize the distinction in Figure 1. Let $\mathcal{C}$ be a finite claim set, $\mathcal{S}$ scientific subjects, and $\mathcal{U}$ evidence sources. A subject map $\sigma : \mathcal{C} \to \mathcal{S}$ identifies what each claim is about, and $\preceq$ orders claim strength. A contract

$$\begin{aligned} \mathcal{K} &= (\mathcal{C}, \sigma, \preceq, \Omega, \mathcal{O}, \mathcal{A}, \mathcal{B}, \Pi), \\ c_1 \preceq c_2 &\Rightarrow \big[\sigma(c_1) = \sigma(c_2),\ \mathcal{O}(c_1) \subseteq \mathcal{O}(c_2)\big], \\ d \in \mathcal{B}(c, s) &\Rightarrow d \parallel c. \end{aligned} \tag{1}$$

contains evidence-obligation types Ω, claim requirements $\mathcal{O}(c)$, the downward-closed claims $\mathcal{A}(s)$ admissible from source s, retained-credit candidates $\mathcal{B}(c, s)$, and an outcome policy Π. Thus weakening stays within a scientific subject, whereas $\mathcal{B}$ preserves licensed conclusions outside the requested weakening branch; when $\sigma(d) \neq \sigma(c)$, such credit is cross-subject.

For evidence record $E = (s, T, q, M)$ (source, telemetry, comparator, and adjudication record), each obligation resolves to PASS, FAIL, MISSING, or UNRESOLVED; the main policy licenses only PASS. Writing $\bar{o}_\Pi(E) \in \{0, 1\}$ for this resolved outcome, define

$$\begin{aligned} \mathcal{L}_\mathcal{K}(E) &= \{c \in \mathcal{A}(s) : \bar{o}_\Pi(E) = 1\ \forall o \in \mathcal{O}(c)\}, \\ \mathcal{C}^\star &= \mathrm{Max}_\preceq\{c \in \mathcal{L}_\mathcal{K}(E) : c \preceq c_{\mathrm{req}},\ \ \sigma(c) = \sigma(c_{\mathrm{req}})\}, \\ \mathcal{R}^\star &= \mathrm{Max}_\preceq\big(\mathcal{L}_\mathcal{K}(E) \cap \mathcal{B}(c_{\mathrm{req}}, s)\big). \end{aligned} \tag{2}$$

$\mathcal{C}^\star$ answers the requested scientific question; $\mathcal{R}^\star$ preserves positive credit outside that branch. The requested status is SUPPORTED if $c_{\mathrm{req}} \in \mathcal{C}^\star$, NARROWED if $\mathcal{C}^\star \neq \varnothing$ but excludes c_{req}, and UNSUPPORTED otherwise. A failure record F stores source mismatch and nonpassing requested-claim obligations, giving output $\Gamma_\mathcal{K}(c_{\mathrm{req}}, E) = (\mathcal{C}^\star, \mathcal{R}^\star, F, z)$. Crucially, z depends only on $\mathcal{C}^\star$.

Two properties clarify the representation. First, identical resolved evidence can license different scopes under different coherent contracts because the contract determines how evidence requirements attach to claims. Second, licensed retained credit is disjoint from the requested weakening branch and therefore cannot substitute for requested-subject support. Appendix C.5 gives the constructions and proofs, together with downward closure, maximal coverage, and monotonicity under licensing-favorable evidence refinement.

4 ROUTING INSTANTIATION

4.1 ROUTING EVIDENCE AND ATTRIBUTION

We instantiate the semantics for preference-conditioned routing on directed multigraphs. At time t, a router receives a graph

$$G_t = (V_t, E_t)$$

and request

$$r_t = (s_t, d_t, x_t, \phi_t, H_t),$$

where s_t, d_t are terminals, x_t is context, $\phi_t \in \Delta^{K-1}$ is a preference vector over K objectives, and H_t denotes hard constraints. Parallel actions retain edge identity $e = (u, v, b)$. With vector edge costs $j_e(t, r_t)$, path cost is

$$J(P; r_t) = \sum_{e \in P} j_e(t, r_t).$$

Let $\mathcal{F}(G_t, r_t)$ be the feasible path set. The exact oracle is

$$P_t^\star(\phi_t) = \arg \min_{P \in \mathcal{F}(G_t, r_t)} \phi_t^\top J(P; r_t),$$

and regret of returned path $\widehat{P}_t$ is

$$\mathrm{Regret}(\widehat{P}_t) = \phi_t^\top J(\widehat{P}_t; r_t) - \phi_t^\top J(P_t^\star; r_t). \tag{3}$$

Regret measures downstream decision quality, but it does not determine which scientific subject produced that quality. Low end-to-end regret may arise from the learned primary policy, repair, solver delegation, or their combination. Claim scope therefore additionally depends on attribution, task identity, preference coverage, dynamics, source scope, and provenance.

4.2 LEARNED-PRIMARY CLAIM BRANCH

The requested learned-primary family is the cumulative chain

$$\mathcal{C}_{\mathrm{lp}} = \{C0 \prec C1 \prec \cdots \prec C6\},$$

with $\sigma(C_i) = \textsf{learned-primary}$ for every $C_i \in \mathcal{C}_{\mathrm{lp}}$.

The complete contract also contains typed families for backup-enabled hybrid conclusions, source-native or bridge conclusions, and diagnostic conclusions:

$$\mathcal{C} = \mathcal{C}_{\mathrm{lp}} \cup \mathcal{C}_{\mathrm{hyb}} \cup \mathcal{C}_{\mathrm{src}} \cup \mathcal{C}_{\mathrm{diag}}. \tag{4}$$

Claims in these other families may be retained in $\mathcal{R}^\star$, but they do not lie on the learned-primary weakening branch.

For a requested learned-primary level L_c and source-supported cap L_s,

$$L^\star = \max \{\ell \le \min(L_c, L_s) : C_\ell \in \mathcal{L}_{\mathcal{K}}(E)\}, \tag{5}$$

with $L^\star = \bot$ when no learned-primary level is licensed. In this chain, $\mathcal{C}^\star$ is therefore empty or a singleton, while $\mathcal{R}^\star$ may contain positive conclusions from the hybrid, source-native, or diagnostic families.

4.3 WHEN GOOD SYSTEM PERFORMANCE CHANGES SUBJECT

Fallback provides a direct example. Consider a request for a $C5$ conclusion about a standalone learned-primary router. One controlled policy obtains the lowest end-to-end regret in the routing suite, but an exact backup solver rescues 86.1% of requests. The observed aggregate performance therefore does not provide the primary-policy attribution required by the learned branch.

For this record, $\mathcal{C}^\star = \varnothing$ for the requested learned-primary claim. The same evidence is nevertheless informative about the executed system: end-to-end feasibility and safety-control conclusions are licensed for the backup-enabled hybrid and retained in $\mathcal{R}^\star$.

Table 1: Learned-primary claim branch used in the routing studies. Each level inherits lower-level requirements; primary-policy attribution and adjudication-record closure are branch-common.

Level	Licensed learned-primary conclusion	Incremental evidence
$C0$	Scalar decision quality	Primary-only or intervention-free regret against a declared comparator.
$C1$	Feasibility-audited routing	Exact feasibility and invalid-route accounting.
$C2$	Identity-preserving multigraph routing	Edge identity, bundle-aware actions, and collapse checks.
$C3$	Preference-conditioned routing	Preference sampling, coverage/hypervolume, and held-out preference shift.
$C4$	Intervention-audited constrained routing	Constraint masks, fallback/repair taxonomy, and primary versus end-to-end decomposition.
$C5$	Dynamic robust routing	Drift stress, route-change evidence, tail regret, and latency.
$C6$	Deployment-style support	Representative workloads, mechanism diagnosis, intended-use limits, and independent reproduction or usability.

Thus the adjudication records two facts simultaneously: the evidence does not positively answer the requested standalone learned-policy question, and it does support a positive conclusion about the larger system. A branch-only weakening output would retain the first fact but not the second.

Claim Cards serialize this decomposition by recording the requested subject and claim, source, comparator, observed evidence, licensed requested conclusion, retained credit, typed failures, and provenance. Appendix C gives the full schemas, executable rules, and representative cards.

5 Evaluation Design

We evaluate subject-typed licensing along the path from scientific text to scientific decision. The studies address three questions.

RQ1: Reconstructability and execution. Can non-authors recover the claim–evidence objects required by the contract and apply the declared semantics consistently?

RQ2: Reviewer judgments. How does a shared subject-typed semantics change review judgments relative to unstructured careful review and a structured evidence checklist?

RQ3: Decision consequences. Can score-only ranking and selection conditioned on a requested scientific claim lead to different decisions over the same evidence records?

5.1 Reconstructing and executing the contract

The first study separates *object reconstruction* from *contract execution*.

For object reconstruction, two non-author coders independently inspect prespecified sections associated with 24 claims and recover comparator identity, source family, requested claim family, intended scope, and required evidence obligations. Coders are blind to the author encodings and adjudication outputs. Agreement therefore measures whether the objects consumed by the licensing procedure can be reconstructed from the underlying scientific text.

The fixed-contract study holds the evidence packets and contract constant. Four blinded annotators apply the learned-primary semantics to 120 claims in ClaimScope-120. We measure exact agreement, ordinal agreement on the $C0$–$C6$ chain, agreement over evidence requirements, and agreement between executable adjudication and the contract-relative human majority. This separates reproducibility of the input representation from consistency of the declared inference procedure.

5.2 HELD-OUT REVIEWER COMPARISON

The reviewer comparison uses 40 claims from 20 additional public papers that did not contribute to construction of the learned-primary branch, its evidence requirements, examples, or CLAIMSCOPE-120. Eight non-author experts evaluate the same prespecified abstract, method, experiment, results, and limitations sections under three conditions.

Careful review. Reviewers determine the strongest supported conclusion without a structured aid.

Evidence checklist. Reviewers receive structured prompts for comparator choice, feasibility, multi-graph identity, preference coverage, fallback and attribution, dynamics, latency, source scope, artifact closure, and wording ambiguity, while writing the final conclusion themselves.

SCOPE-ROUTING. Reviewers receive the same evidence fields together with the declared learned-primary inheritance relation and maximal-selection semantics.

Each claim receives two independent judgments per condition, for 240 judgments in total, using a balanced incomplete-block assignment in which no reviewer encounters the same claim under multiple conditions. A prespecified pair-resolution procedure produces one claim-level summary for each condition while retaining both raw judgments, rationales, disagreement records, and timing fields.

Separately, three senior experts, blind to the learned-primary ladder and condition outputs, write in free form the strongest conclusion they consider supported by the supplied evidence. The panel resolves 36 of the 40 claims. Their consensus statements are mapped through the prespecified resolution procedure to form a common operational reference for the three reviewer conditions.

We report agreement with this reference together with reference-relative upward and downward deviations. Comparisons are paired by claim, and uncertainty intervals resample papers as clusters. Full prompts, assignments, resolution procedures, and reliability analyses appear in Appendix E.5.

5.3 CONTROLLED ROUTING CONSEQUENCES

ROUTESCOPE-LITE studies how claim semantics interact with selection under a common evidence schema. The suite contains 1,500 requests spanning three semisynthetic domain families, 15 graph-family slices, 10 seeds, and 15 policy families. The policy inventory includes learned policies, classical comparators, negative controls, repair and fallback systems, preference-conditioned models, and dynamic policies that vary in attribution, multigraph identity, preference coverage, dynamic response, and latency.

We prespecify $C5$, dynamic robustness of the learned-primary router, as the target claim. Two decision rules are applied to the same learned-method rows. *Score-only selection* ranks methods by aggregate regret. *Claim-conditioned selection* first restricts attention to methods licensed for the target claim under the declared contract and then applies the prespecified quality rule within that eligible set.

This design separates two scientific questions: which method minimizes an observed score, and which method is supported for a particular scientific conclusion. Appendices F–G provide the complete policy inventory, uncertainty analyses, source-scope audits, obligation-removal analyses, and alternative missing-evidence policies.

6 RESULTS

6.1 RQ1: RECONSTRUCTABILITY AND EXECUTION

Object reconstruction. The two non-author coders independently recover comparator identity for 23/24 claims, source family for 22/24, and requested claim family for 21/24. Intended-scope agreement reaches weighted $\kappa = 0.78$, and mean Jaccard agreement over evidence requirements is 0.82. Independently reconstructed final status agrees on 20/24 claims, and all status disagreements are within one learned-primary level.

Table 2: Held-out reviewer comparison on 36 expert-resolved claims.

Outcome	Careful review	Evidence checklist	SCOPE-ROUTING
Agreement with operational reference	24/36	27/36	31/36
Reference-relative upward deviation	8/36	5/36	2/36
Reference-relative downward deviation	1/36	3/36	4/36
Median time per claim	18.1 min	13.4 min	15.8 min

The required objects are therefore largely recoverable from the prespecified paper sections, with the remaining disagreements concentrated near scope boundaries.

Fixed-contract execution. On CLAIMSCOPE-120, four blinded annotators achieve mean pairwise exact agreement 0.761, mean weighted $\kappa = 0.703$, and ordinal Krippendorff's $\alpha = 0.686$. SCOPE-ROUTING matches the contract-relative human majority exactly on 101/120 claims and within one level on 115/120.

The remaining exact disagreements occur primarily near adjacent coverage, attribution, dynamics, and latency boundaries. Appendix E reports complete reliability statistics and tie-policy sensitivity.

6.2 RQ2: REVIEWER OPERATING POINTS

Table 2 reports the 36 claims for which the independent expert panel produced a resolved operational reference.

Relative to careful review, the paired paper-clustered agreement difference for SCOPE-ROUTING is 19.4 percentage points with a 95% interval of $[5.6, 33.3]$. Relative to the evidence checklist, the observed difference is 11.1 percentage points with interval $[-2.8, 25.0]$.

The directional results characterize a different observed review operating point. Upward deviations decrease from 8/36 under careful review to 5/36 under the checklist and 2/36 under SCOPE-ROUTING. Downward deviations increase from 1/36 to 3/36 and 4/36, respectively. Median review time is highest for careful review, lowest for the checklist, and intermediate for SCOPE-ROUTING.

To characterize what changes rather than only how often judgments differ, we inspect the 12/40 held-out claims for which subject-typed licensing changes the initial score-oriented interpretation. Eleven have resolved expert references: nine changes align with the reference and two are judged more conservative. The endorsed cases include loss of multigraph identity, missing preference-space coverage, and absence of dynamic-response evidence. Among eight narrowed cases, five retain positive lower-scope credit rather than collapsing to a purely negative conclusion.

6.3 RQ3: CLAIM-CONDITIONED DECISIONS

The controlled routing suite produces a direct separation between score minimization and target-claim eligibility.

Among learned policy families, *Fallback-Heavy Safe* attains the lowest aggregate regret, 0.011091, and is therefore selected by the regret-only rule. Its end-to-end performance, however, includes exact backup rescue on 86.1% of requests. The learned-primary attribution required by the target $C5$ claim is therefore not licensed. Its positive hybrid-system feasibility and safety-control evidence is retained in $\mathcal{R}^\star$.

Shock-Safe Online Repair has higher aggregate regret, 0.016189, but satisfies the inherited evidence requirements for the target $C5$ learned-primary claim and is selected by the target-$C5$-then-regret rule.

Table 3: Selection depends on the scientific decision rule.

Decision rule	Selected row	Regret	Key evidence	Licensed interpretation
Score only	Fallback-Heavy Safe	0.011091	86.1% backup rescue	Hybrid credit retained; no learned-primary $C5$.
Target $C5$, then regret	Shock-Safe Online Repair	0.016189	Dynamic and attribution evidence complete	$C5$ learned-primary.
Coverage-oriented diagnostic	Preference Hypernetwork	0.020865	HV $= 0.849$	$C4$; $C5$ dynamics incomplete.

Preference Hypernetwork has aggregate regret 0.020865 and stronger preference-space coverage, with hypervolume 0.849, but remains at $C4$ because the $C5$ dynamic requirements are incomplete.

Thus the selected method depends on the scientific question being asked. Score-only selection asks which row minimizes aggregate regret. Claim-conditioned selection asks which rows first provide evidence for the requested learned-primary conclusion and only then compares quality within that eligible set.

Across the shared learned-method set, the score-based and claim-conditioned orderings have Kendall $\tau_b = 0.407$. Leave-one-obligation-out analysis additionally produces distinct conclusion changes associated with fallback attribution, multigraph identity and feasibility, preference coverage, route-change evidence, latency, source scope, and record closure. The principal rescue-attribution, source-transfer, and deployment-scope boundaries remain unchanged across the three declared missing-evidence policies. Appendix G reports the complete analyses.

Across the three evaluation stages, the same semantic distinction appears at different levels. Non-authors can largely reconstruct and execute the declared claim objects; shared semantics changes which conclusions reviewers license from the same evidence; and the controlled routing suite shows that a score-optimal row need not be eligible for a requested component-level claim even when it retains positive system-level credit. The common object across these results is not a global method ranking, but the relation between an evidence record, a scientific subject, and the conclusion that the record supports.

7 Implications and Scope

7.1 Claim strength and scientific subject

Subject-typed licensing separates two questions that are often collapsed in benchmark interpretation: *what can the evidence support about the requested scientific subject, and what other positive conclusions does the same evidence support?* The first question is answered by requested-subject licensing; the second by retained credit. Their separation allows evidence to remain informative without making it substitutable across components or system boundaries.

This distinction matters whenever reported performance is produced by a composition of mechanisms. A learned component may operate with repair, search, external tools, retrieval, or safety controls. In such systems, end-to-end behavior can be strong even when the evidence required to attribute that behavior to one component is incomplete. Subject-typed reporting preserves both observations: the component-level conclusion remains matched to its evidence, while positive system-level behavior remains visible.

A benchmark row can therefore remain scientifically positive while changing subject. Rather than pairing a method name with a scalar score alone, an evidence-qualified result connects a measured outcome to the subject to which that outcome is attributed, the evidence required for that conclusion, and the scope from which the evidence was obtained.

7.2 CLAIM CONTRACTS AS SCIENTIFIC OBJECTS

A claim contract is *normative, versioned, and contestable*. Its subject assignments, claim-strength relations, source rules, and evidence requirements encode a scientific position about what observations are sufficient for particular conclusions. Making these choices explicit does not remove scientific judgment; it makes part of that judgment inspectable.

This representation also localizes disagreement. Two evaluators may agree on the observed telemetry yet differ because they adopt different source boundaries, evidence requirements, or claim relations. Such disagreement is different from disagreement about the underlying measurements. A versioned contract makes the distinction visible and allows changes in scientific scope to be traced to explicit changes in claim semantics.

This perspective suggests a broader role for claim contracts in evaluation practice. Benchmark designers can expose the evidence required for intended interpretations; authors can associate reported results with the subject and scope they support; and reviewers can inspect the relation between observations and conclusions rather than reconstructing that relation implicitly for every paper.

7.3 EMPIRICAL SCOPE

The empirical studies establish three results for the declared learned-routing semantics. First, non-author coders and annotators can recover and execute the required claim–evidence objects with substantial, though imperfect, agreement. Second, the held-out comparison shows that supplying shared claim semantics changes the observed review operating point relative to both unstructured review and a structured evidence checklist. Third, the controlled routing suite provides a mechanistic example in which score-only and target-claim decisions diverge while positive credit for the executed hybrid system is preserved.

These studies operate at different levels of evidence. CLAIMSCOPE-120 and the construction audit study reproducibility of interpretation from supplied paper sections; the reviewer experiment studies judgments on held-out public-paper claims; and ROUTESCOPE-LITE studies controlled consequences under semisynthetic routing conditions. Public source cases additionally illustrate how native, bridge, diagnostic, and out-of-contract conclusions interact with source scope.

The resulting empirical picture is therefore one of *representation, reproducibility, and controlled consequence*, rather than prevalence. The present studies do not estimate how often cross-subject attribution occurs throughout the ML literature, and the semisynthetic routing records do not constitute deployment evidence. Broader task sampling, independently developed contracts, and operational data would address those questions at a different empirical level.

8 CONCLUSION

Partial support in a composite learned system has more than one direction. Evidence can justify a weaker conclusion about the requested scientific subject, or it can justify a positive conclusion about a different component or system. Subject-typed claim licensing represents these outcomes separately through requested-subject support and non-substitutive retained credit. The routing studies show that this distinction can be reconstructed and executed by non-authors, can alter the operating point of scientific review, and can separate score minimization from eligibility for a particular scientific claim. More broadly, the framework treats claim strength and scientific subject as distinct coordinates of evidence-based interpretation. Weakening answers what can still be said about the requested subject; retained credit records what else the evidence positively supports. Keeping both visible—without allowing one to answer for the other—is the central purpose of subject-typed claim licensing.

AI USE STATEMENT

Generative AI tools were used for literature discovery and editorial assistance with the wording, organization, and presentation of the manuscript. The authors independently read and verified the cited sources, developed and checked the formal framework, designed and conducted the studies, constructed and analyzed the empirical records, and interpreted the results. All AI-assisted text was

reviewed by the authors, who take responsibility for the final manuscript, claims, code, data, and artifacts.

ETHICS STATEMENT

The human-subject component involved adult experts making professional judgments about scientific claims drawn from public papers. The protocol used no deception, collected no sensitive personal information, and releases no identifying annotator data. Participation, confidentiality, compensation, and data handling were disclosed to participants. The protocol received the applicable institutional determination and was classified as exempt; institution-identifying details are withheld during double-blind review. The artifact records anonymized assignments, consent and compensation documentation, conflict-of-interest checks, and the analysis protocol.

REPRODUCIBILITY STATEMENT

The submission artifact contains the adjudicator, claim contracts, schemas, Claim-Card exporters, analysis scripts, anonymized judgment records, locked same-contract rows, manifests, and reviewer commands. Section 5 summarizes the study design; Appendices C, D, E, E.5, F, and I document the executable rules, data construction, statistical analyses, routing suite, and versioned reviewer path. The central ROUTESCOPE-LITE selection result is reanalyzed from locked request–policy rows; the default path distinguishes fresh execution, cached reanalysis, and derived-only historical evidence.

REFERENCES

Mubashara Akhtar, Omar Benjelloun, Costanza Conforti, Luca Foschini, Pieter Gijsbers, Joan Giner-Miguelez, Sujata Goswami, Nitisha Jain, Michalis Karamousadakis, Satyapriya Krishna, et al. Croissant: A metadata format for ml-ready datasets. *Advances in Neural Information Processing Systems*, 37:82133–82148, 2024.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avi Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *International conference on learning representations*, volume 2024, pp. 9112–9141, 2024.

Andrew M Bean, Ryan Othniel Kearns, Angelika Romanou, Franziska Sofia Hafner, Harry Mayne, Jan Batzner, Negar Foroutan Eghlidi, Chris Schmitz, Karolina Korgul, Hunar Batra, et al. Measuring what matters: Construct validity in large language model benchmarks. *Advances in Neural Information Processing Systems*, 38, 2026.

Irwan Bello, Hieu Pham, Quoc V Le, Mohammad Norouzi, and Samy Bengio. Neural combinatorial optimization with reinforcement learning. *arXiv preprint arXiv:1611.09940*, 2016.

Quentin Berthet, Mathieu Blondel, Olivier Teboul, Marco Cuturi, Jean-Philippe Vert, and Francis Bach. Learning with differentiable pertubed optimizers. *Advances in neural information processing systems*, 33:9508–9519, 2020.

Jinbiao Chen, Zhiguang Cao, Jiahai Wang, Yaoxin Wu, Hanzhang Qin, Zizhen Zhang, and Yue-Jiao Gong. Rethinking neural multi-objective combinatorial optimization via neat weight embedding. In *International Conference on Learning Representations*, volume 2025, pp. 48510–48529, 2025.

Kalyanmoy Deb, Amrit Pratap, Sameer Agarwal, and TAMT Meyarivan. A fast and elitist multiobjective genetic algorithm: Nsga-ii. *IEEE transactions on evolutionary computation*, 6(2):182–197, 2002.

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A Smith. Show your work: Improved reporting of experimental results. 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)*, pp. 2185–2194, 2019.

Adam N Elmachtoub and Paul Grigas. Smart "predict, then optimize". *Management science*, 68(1): 9–26, 2022.

Shengyu Feng, Weiwei Sun, Shanda Li, Ameet Talwalkar, and Yiming Yang. Frontierco: Real-world and large-scale evaluation of machine learning solvers for combinatorial optimization. In *International Conference on Learning Representations*, volume 2026, pp. 118924–118948, 2026.

Timo Freiesleben and Sebastian Zezulka. The benchmarking epistemology: Construct validity for evaluating machine learning models. *arXiv preprint arXiv:2510.23191*, 2025.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92, 2021.

Marah Ghoummaid and Uri Shalit. When to act and when to ask: policy learning with deferral under hidden confounding. *Advances in Neural Information Processing Systems*, 37:56108–56135, 2024.

Ahmed Heakl, Yahia Salaheldin Shaaban, Salem Lahlou, Martin Takac, and Zangir Iklassov. Svrp-bench: A realistic benchmark for stochastic vehicle routing problem. *Advances in Neural Information Processing Systems*, 38, 2026.

Joseph James, Chenghao Xiao, Yucheng Li, Nafise Sadat Moosavi, and Chenghua Lin. Rigourate: Quantifying scientific exaggeration with evidence-aligned claim evaluation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 34022–34043, 2026.

Peter Jansen, Samiah Hassan, and Ruoyao Wang. Matter-of-fact: A benchmark for verifying the feasibility of literature-supported claims in materials science. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 4090–4102, 2025.

Shashidhar Reddy Javaji, Yupeng Cao, Haohang Li, Yangyang Yu, Nikhil Muralidhar, and Zining Zhu. Can ai validate science? benchmarking llms on claim→ evidence reasoning in ai papers. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 2355–2379, 2025.

Michael T Kane. Validating the interpretations and uses of test scores. *Journal of educational measurement*, 50(1):1–73, 2013.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, et al. Dynabench: Rethinking benchmarking in nlp. In *Proceedings of the 2021 conference of the North American chapter of the Association for Computational Linguistics: human language technologies*, pp. 4110–4124, 2021.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International conference on machine learning*, pp. 5637–5664. PMLR, 2021.

Wouter Kool, Herke Van Hoof, and Max Welling. Attention, learn to solve routing problems! *arXiv preprint arXiv:1803.08475*, 2018.

Alan A Lahoud, Erik Schaffernicht, and Johannes A Stork. Datasp: A differential all-to-all shortest path algorithm for learning costs and predicting paths with context. *arXiv preprint arXiv:2405.04923*, 2024.

Hongmin Li. The calibration turn in ai-assisted research: A conceptual and methodological framework for evidence-licensed claims. *arXiv preprint arXiv:2606.31273*, 2026.

Ke Li, Fei Liu, Zhenkun Wang, and Qingfu Zhang. Destroy and repair using hyper-graphs for routing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 18341–18349, 2025.

Sirui Li, Zhongxia Yan, and Cathy Wu. Learning to delegate for large-scale vehicle routing. *Advances in neural information processing systems*, 34:26198–26211, 2021.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.

Jayanta Mandi, James Kotary, Senne Berden, Maxime Mulamba, Victor Bucarey, Tias Guns, and Ferdinando Fioretto. Decision-focused learning: Foundations, state of the art, benchmark and future opportunities. *Journal of Artificial Intelligence Research*, 80:1623–1701, 2024.

Anqi Mao, Mehryar Mohri, and Yutao Zhong. Regression with multi-expert deferral. *arXiv preprint arXiv:2403.19494*, 2024.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 220–229, 2019.

Mathias Niepert, Pasquale Minervini, and Luca Franceschi. Implicit mle: backpropagating through discrete exponential family distributions. *Advances in Neural Information Processing Systems*, 34: 14567–14579, 2021.

Yuxin Pan, Zhiguang Cao, Chengyang Gu, Liu Liu, Peilin Zhao, Yize Chen, and Fangzhen Lin. Multi-task vehicle routing solver via mixture of specialized experts under state-decomposable mdp. *Advances in Neural Information Processing Systems*, 38:32577–32620, 2026.

Felix Petersen, Christian Borgelt, Tobias Sutter, Hilde Kuehne, Oliver Deussen, and Stefano Ermon. Newton losses: Using curvature information for learning with differentiable algorithms. *Advances in Neural Information Processing Systems*, 37:26113–26135, 2024.

Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Larivière, Alina Beygelzimer, Florence d'Alché Buc, Emily Fox, and Hugo Larochelle. Improving reproducibility in machine learning research (a report from the neurips 2019 reproducibility program). *Journal of machine learning research*, 22(164):1–20, 2021.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, et al. Toolllm: Facilitating large language models to master 16000+ real-world apis. In *International Conference on Learning Representations*, volume 2024, pp. 9695–9717, 2024.

Anka Reuel, Amelia Hardy, Chandler Smith, Max Lamparth, Malcolm Hardy, and Mykel J Kochenderfer. Betterbench: Assessing ai benchmarks, uncovering issues, and establishing best practices. *Advances in Neural Information Processing Systems*, 37:21763–21813, 2024.

Thomas Rhodes, Frederick Boland, Elizabeth Fong, and Michael Kass. Software assurance using structured assurance case models. *Journal of research of the National Institute of Standards and Technology*, 115(3):209, 2010.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of nlp models with checklist. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 4902–4912, 2020.

Filip Rydin, Attila Lischka, Jiaming Wu, Morteza Haghir Chehreghani, and Balázs Kulcsár. Beyond simple graphs: Neural multi-objective routing on multigraphs. In *International Conference on Learning Representations*, volume 2026, pp. 78021–78054, 2026.

Olawale Salaudeen, Anka Reuel, Ahmed Ahmed, Suhana Bedi, Zachary Robertson, Sudharsan Sundar, Ben Domingue, Angelina Wang, and Sanmi Koyejo. Measurement to meaning: A validity-centered framework for ai evaluation. *arXiv preprint arXiv:2505.10573*, 2025.

Anna Sokol, Elizabeth Daly, Michael Hind, David Piorkowski, Xiangliang Zhang, Nuno Moniz, and Nitesh Chawla. Benchmarkcards: Standardized documentation for large language model benchmarks. *Advances in Neural Information Processing Systems*, 38, 2026.

Weiwei Sun, Shengyu Feng, Shanda Li, and Yiming Yang. Co-bench: Benchmarking language model agents in algorithm search for combinatorial optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 33126–33134, 2026.

Bo Tang and Elias Boutros Khalil. Pyepo: A pytorch-based end-to-end predict-then-optimize library with linear objective function. In *OPT 2022: Optimization for Machine Learning (NeurIPS 2022 Workshop)*, 2022.

Liyan Tang, Philippe Laban, and Greg Durrett. Minicheck: Efficient fact-checking of llms on grounding documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8818–8847, 2024.

Bryan Wilder, Bistra Dilkina, and Milind Tambe. Melding the data-decisions pipeline: Decision-focused learning for combinatorial optimization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 1658–1665, 2019.

Pei Xiao, Zizhen Zhang, Jinbiao Chen, Jiahai Wang, and Zhenzhen Zhang. Neural combinatorial optimization for robust routing problem with uncertain travel times. *Advances in Neural Information Processing Systems*, 37:134841–134867, 2024.

Ling Yue, Chaoqian Ouyang, Hang Xu, Ruijun Huang, Yuchen Liu, Libin Zheng, Wei Liu, Shaowu Pan, Shimin Di, and Min-Ling Zhang. Factreview: Evidence-grounded peer review with execution-based claim verification. *arXiv preprint arXiv:2604.04074*, 2026.

Jianan Zhou, Yaoxin Wu, Zhiguang Cao, Wen Song, Jie Zhang, and Zhiqi Shen. Collaboration! towards robust neural methods for routing problems. *Advances in Neural Information Processing Systems*, 37:121731–121764, 2024.

A Extended Positioning: Feature-Level Comparisons and Boundaries

Section 2 traces prior work by its primary output along the path from measurement to reportable claim. This appendix provides the complementary feature-level view: it records the specific representational boundaries, composition roles, and adjudication outputs of adjacent approaches.

Argument-based validity and AI evaluation. Argument-based validity treats the interpretations and uses inferred from scores as the object to be justified Kane (2013). Construct-validity and validity-centered AI evaluation frameworks apply the same principle to benchmark claims Freiesleben & Zezulka (2025); Salaudeen et al. (2025); Bean et al. (2026). SCOPE-Routing supplies an executable representation for one bounded decision in that process: a row-level requested claim, a declared strength order and source-admissibility relation, claim-indexed obligations, maximal supported elements, typed failures, retained incomparable credit, and an artifact-bound rationale.

Evidence-licensed calibration and claim replay. The calibration turn frames evidence-licensed assertion as a scientific operation and maps raw claims to maximal licensed frontiers Li (2026). Commit-bound claim replay makes the inference licensed by evaluation evidence executable and studies stability across evaluator-meaning families. Composite systems introduce two roles for licensed conclusions: answering the requested subject and preserving non-substitutive credit about another subject. The SCOPE contract represents these roles through $\mathcal{C}^\star$ and $\mathcal{R}^\star$, with source admissibility and component-level telemetry making the attribution decision executable. The routing studies evaluate reconstruction, review behavior, and selection consequences for this typed representation.

Structured assurance and claim verification. Structured assurance cases connect claims, subclaims, arguments, evidence, assumptions, and context in traceable graphs and permit top-down, bottom-up, or mixed construction Rhodes et al. (2010). Their support edges describe an argument for a designated claim. CLAIM-BENCH evaluates claim-to-evidence reasoning; RIGOURATE retrieves paper-internal multimodal evidence and predicts continuous overstatement scores with justifications; FactReview combines literature positioning with execution-based verification Javaji et al. (2025); James et al. (2026); Yue et al. (2026). These outputs can populate a contract-defined obligation evaluator. SCOPE-Routing represents the resulting benchmark conclusion through maximal requested-subject reformulations in $\mathcal{C}^\star$ and contract-licensed positive credit about other task or system subjects in $\mathcal{R}^\star$.

Decision-focused learning and ML for combinatorial optimization. Decision-focused learning and predict-then-optimize methods train predictive components through downstream decision loss rather than prediction error alone Elmachtoub & Grigas (2022); Wilder et al. (2019); Mandi et al. (2024). Differentiable optimizers, perturbation-based surrogates, implicit layers, and solver-aware learning have made decision-quality evaluation increasingly common Berthet et al. (2020); Niepert et al. (2021); Petersen et al. (2024). These methods improve how decisions are learned and how decision quality is measured. A low-regret row may reflect learned scoring, fallback rescue, repair, collapse, static equivalence, or task mismatch. SCOPE-ROUTING treats downstream decision quality as one evidence channel and adjudicates the claim scope licensed by the full record.

Neural routing and multigraph routing. Neural routing has produced learned heuristics for classical routing tasks Bello et al. (2016); Kool et al. (2018). Multiobjective and multigraph routing make action identity, preference conditioning, and constraint structure part of the evaluated object Deb et al. (2002); Rydin et al. (2026). These settings expose the score-to-claim gap especially sharply: scalar wins can coexist with lost parallel actions, degraded preference coverage, hidden rescue behavior, or static-equivalent route choices. The routing contribution evaluates the scientific scope licensed by such evidence.

Documentation and evaluation-science artifacts. Datasheets, model cards, BenchmarkCards, holistic evaluation, behavioral testing, dynamic benchmarks, and distribution-shift benchmarks make assumptions, intended uses, coverage, and failure modes explicit Gebru et al. (2021); Mitchell et al. (2019); Sokol et al. (2026); Liang et al. (2022); Ribeiro et al. (2020); Kiela et al. (2021); Koh et al. (2021). These artifacts improve what is reported about data, models, and benchmarks. SCOPE-ROUTING consumes these fields in a row-level executable record containing the reported score, source, telemetry, comparator, adjudication state, and subject-typed conclusions.

Artifact-centered reproducibility. Reproducibility programs and metadata standards emphasize that code, data, schemas, and reporting details are part of scientific evidence Pineau et al. (2021); Dodge et al. (2019); Akhtar et al. (2024). SCOPE-ROUTING requires a locked record for adjudication at every level, while C6 additionally requires independent reproduction and evidence about usability in the intended setting.

Expanded comparison. Table 4 catalogs what each adjacent line contributes to the licensing pipeline. Table 5 isolates the central output-space distinction by comparing the information represented by alternative adjudication strategies.

Table 4: Feature-level positioning of adjacent evaluation artifacts in the claim-licensing pipeline.

Prior line of work	Primary output	Composition role in claim licensing
Argument-based and AI-evaluation validity	Evidence-backed interpretations, uses, and claims.	Supplies validity arguments; the contract operationalizes one bounded row-level conclusion decision.
Evidence-licensed calibration and claim replay	Calibrated assertion scope, maximal licensed frontiers, replayable inference, and semantic stability.	Supplies evidence-relative claim semantics; subject-typed licensing separates requested-subject answers from non-substitutive credit.
Structured assurance cases	Traceable claim–argument–evidence graph.	Supplies explicit support structure that can be encoded as claim obligations.
Claim-evidence retrieval and assessment	Evidence links, support labels, justifications, or graded overstatement scores.	Supplies evidence and support judgments consumed by obligation evaluators.
Datasheets and model cards	Documentation of data, models, intended uses, and limitations.	Supplies inspectable evidence conditions and intended-scope fields.
BenchmarkCards and benchmark documentation	Standardized benchmark metadata and risk reporting.	Supplies benchmark-level evidence records and scope declarations.
Decision-focused evaluation	Downstream regret or decision quality.	Supplies the quality channel; broader claims consume additional obligations.
ML4CO and routing benchmarks	Native task metrics, solver comparisons, and efficiency reports.	Supplies source-native evidence governed by source-admissibility rules.

Table 5: Adjudication strategies compared by represented information and primary output.

Adjudicator	Evidence diagnosis	Claim order	Source semantics	Incomparable credit	Primary output
Regret-only ranking	Scalar score	Absent	Absent	Absent	Score-ordered row.
Missing-field checklist	Structured gap prompts	Reviewer-authored	Prompted or implicit	Reviewer-authored	Gap record plus reviewer-selected claim.
Designated-claim assessment	Evidence links or support scores	Usually absent	Source-bound or implicit	Usually absent	Support label, score, or justification.
Unstructured human review	Contextual	Implicit	Implicit	Implicit	Review prose and judgment.
SCOPE-ROUTING	Typed obligations	Explicit	Explicit	Explicit $\mathcal{B}$ relation	$(\mathcal{C}^\star, \mathcal{R}^\star, F, z)$.

B ARCHITECTURAL PORTABILITY

The reusable object in SCOPE-ROUTING is the claim-licensing architecture. A port to predict-then-optimize selection retains a claim partial order, claim-indexed obligations, source admissibility, the maximal-element operator, observable records, and provenance-bound rationales, with selection-specific claims and gates replacing the routing learned-primary branch Elmachtoub & Grigas (2022); Wilder et al. (2019); Mandi et al. (2024).

Table 6: Worked design sketch for a predict-then-optimize selection port; empirical cross-domain validation remains a separate study.

Architectural role	Routing instantiation	Selection-task replacement
Decision quality	Route regret	Downstream selected-set objective or regret
Feasibility	Valid source–destination path	Budget, cardinality, or packing feasibility
Action identity	Parallel-edge identity	Item identity and duplicate/substitution handling
Coverage	Preference-space hypervolume	Coverage across budgets, preferences, or item strata
Attribution	Fallback solver or repair	Optimizer, repair, abstention, or fallback disclosure
Robustness	Drift, route change, tail regret	Distribution shift, set change, tail decision loss
Systems evidence	Decode and fallback latency	Prediction, optimizer, repair, and end-to-end latency
Deployment	Live routing workload and use limits	Intended-use population, operational constraints, independent use

The port requires a new evidence contract and validation study. Its claim levels and obligations would be authored for the selection task and evaluated with task-specific evidence.

C ADJUDICATOR DETAILS

This appendix gives the implementation-level semantics behind Section 3. The main text presents the row-level computation; here we record the telemetry channels, source-scope defaults, support semantics, and Claim-Card schema used by the artifact.

C.1 HARD AND SOFT EVIDENTIARY BOUNDARIES

The strongest boundary results depend on hard distinctions. Missing multigraph identity blocks C2+ claims; source mismatch blocks unsupported transfer; and an unlocked adjudication record blocks an inspectable SCOPE-ROUTING conclusion. Fallback is treated through attribution: hybrid-system credit may survive, but standalone learned-component credit requires fallback-separated evidence. Softer requirements, such as partial coverage or latency reporting, can be configured more permissively and are reported in sensitivity analyses.

The retained families complete the subject-typed claim universe. $\mathcal{C}_{\mathrm{hyb}}$ contains end-to-end feasibility and safety-control conclusions for the backup-enabled stack; $\mathcal{C}_{\mathrm{src}}$ contains native-task and bridge conclusions; and $\mathcal{C}_{\mathrm{diag}}$ contains mechanism and failure-diagnosis conclusions. Each retained Claim Card declares its subject, source, and obligations, and $\mathcal{B}$ selects only candidates incomparable with the requested learned-primary claim.

C.2 TELEMETRY CHANNELS

Table 7 lists the telemetry channels used by the adjudicator. These channels make the ladder observable: missing edge identity blocks C2, missing preference coverage blocks C3, missing fallback evidence blocks C4, and missing route-change evidence blocks C5.

Table 7: Telemetry channels used by the adjudicator.

Channel	Representative fields	Claim-validity role
Quality	scalarized regret, tail regret, oracle objective, comparator identity	Establishes C0 scoring evidence.
Feasibility	aggregate success, feasible-set-nonempty flag, conditional invalid-route count, hidden-repair flag	Determines whether C1+ feasibility-audited claims can be credited; rounded aggregate rates are not the gate.
Fallback	claim subject, fallback reason/rate/trigger, primary-only and end-to-end quality, fallback-credit flag	Separates learned-component evidence from hybrid rescue and safety-control credit.
Structure	edge identity, bundle id, retained-edge set, collapse-ablation id	Prevents simple-graph evaluation from certifying C2 multigraph claims.
Preference	preference vector, sampler, hypervolume, held-out shift split	Tests whether scalar gains preserve C3 preference-space behavior.
Dynamics	route identity, route-change W/T/L, drift severity, non-static-equivalence flag	Distinguishes dynamic learned behavior from static-equivalent rows.
Systems	scoring latency, decode latency, fallback latency, online-step latency	Blocks operational claims with impractical or hidden latency cost.
Provenance	manifest key, schema version, seed, split id, locked summary anchor	Binds each row to the artifact used for adjudication.

C.3 EVIDENCE-SOURCE SCOPE

Source scope is applied before row-level telemetry gates. A row can support only claims whose required structures are instantiated by the source that produced the evidence. Table 8 gives the default

source-scope rule. This is the firewall that allows SCOPE-ROUTING to credit public evidence at its native or bridge scope without converting it into same-contract dynamic preference-conditioned multigraph-routing evidence.

Table 8: Evidence-source scope.

Evidence source	Instantiated structure	Default admissible claim scope
Controlled multigraph generator	Score, edge identity, preference vectors, hard feasibility, and controlled drift.	Controlled C1–C5 claims under generator assumptions.
Same-contract ROUTESCOPE-LITE suite	Multidomain semisynthetic graphs, shared routing contract, 15 policy families, 10 seeds, telemetry-complete rows.	C1–C5 same-contract evidence; C6 additionally requires deployment evidence.
Public shortest-path bridges	Score transfer under a public shortest-path or contextual path contract.	C0 learned-scoring bridge credit; routing C3 additionally requires inherited feasibility and multigraph-identity obligations.
Diagnostic realistic slices	Realistic topology or dynamics with limited slice size and mechanism probes.	Mechanism diagnosis unless the slice is large enough and telemetry-complete enough for confirmatory realistic claims.
Native or external ML4CO adapters	Task-dependent objectives, metrics, and solver-interface behavior.	Compatibility, native-metric, efficiency, or out-of-contract claims; matching route objects and telemetry enable routing claims.
External public-claim corpus	Four blinded non-author annotation streams and executable labels of claims from published evidence snippets.	Contract-relative operational reference for the specified evidence packets.

C.4 PROMOTION, NARROWING, AND REJECTION

In the general operator, a request is SUPPORTED when it belongs to $\mathcal{C}^\star$, NARROWED when $\mathcal{C}^\star$ is nonempty and excludes the request, and UNSUPPORTED when $\mathcal{C}^\star$ is empty. Native, bridge, diagnostic, and different-subject credit is stored independently in $\mathcal{R}^\star$. The routing implementation uses a chain, so the requested-branch maximal set has at most one C-level. Failure codes distinguish missing evidence, observed failure, source mismatch, and an unlocked adjudication record.

Listing 1 gives the implementation-level rule.

Listing 1: Partial-order claim-licensing rule. The routing implementation uses a chain, so maximal contains at most one C-level.

```
for row in candidate_rows:
   request = requested_claim(row)
   evidence = evidence_record(row)
   contract = declared_claim_contract(row)
   licensed = {c for c in contract.claims
            if source_admissible(evidence.source, c)
            and all(contract.resolve(o, evidence)
                  for o in obligations(c))}
   requested_branch = {c for c in licensed if c <= request}
   maximal = maximal_elements(requested_branch)
   retained = maximal_elements(
      licensed & contract.retained_claims(
         request, evidence.source))
   failures = typed_failures(request, evidence)
```

```
   if request in maximal:
      support(row, claims=maximal, failures=failures,
            retained_credit=retained)
   elif maximal:
      narrow(row, claims=maximal, failures=failures,
           retained_credit=retained)
   else:
      reject(row, failures=failures,
           retained_credit=retained)
```

C.5 Separation and execution properties

The first two constructions establish the separation results stated in Section 3; the remaining properties characterize execution under a fixed declared contract.

Construction for contract dependence. Let $c_0 \prec c_1 \prec c_2$, let $\Omega = \{\ell, p, q\}$ with ℓ the common record-lock obligation, and take one source with $\mathcal{A}(s) = \mathcal{C}$. Fix $c_{\mathrm{req}} = c_2$ and the diagnostic signature

$$\bar{\ell}_\Pi(E) = 1, \qquad \bar{p}_\Pi(E) = 1, \qquad \bar{q}_\Pi(E) = 0.$$

Assign all three claims the same subject and let the two contracts share $(\mathcal{C}, \sigma, \preceq, \Omega, \mathcal{A}, \mathcal{B}, \Pi)$ while differing only in their obligation assignment:

$$\begin{aligned} &\mathcal{O}_1(c_0) = \{\ell\}, \quad \mathcal{O}_1(c_1) = \{\ell, p\}, \quad \mathcal{O}_1(c_2) = \{\ell, p, q\}, \\ &\mathcal{O}_2(c_0) = \{\ell\}, \quad \mathcal{O}_2(c_1) = \{\ell, q\}, \quad \mathcal{O}_2(c_2) = \{\ell, p, q\}. \end{aligned}$$

Both assignments satisfy obligation inheritance. The first licenses $\{c_0, c_1\}$ and returns $\mathcal{C}_1^\star = \{c_1\}$; the second licenses $\{c_0\}$ and returns $\mathcal{C}_2^\star = \{c_0\}$. The requested claim and diagnostic signature coincide, while positive claim scope differs. Hence every universally correct decoder also consumes contract semantics.

Proof of retained-credit separation. Because X_B is a nonempty subset of a finite poset, it has a maximal element and $\mathcal{R}^\star = \mathrm{Max}_{\preceq}(X_B)$ is nonempty. Every element of $\mathcal{B}(c_{\mathrm{req}}, s)$ is incomparable with c_{req}, so no element of $\mathcal{R}^\star$ lies in $\downarrow c_{\mathrm{req}} \cap \mathcal{C}_{\sigma_{\mathrm{req}}}$. An output restricted to requested-subject reformulation therefore omits every element of $\mathcal{R}^\star$. Moreover, every $d \in X_B$ lies below some element of $\mathcal{R}^\star$, giving credit preservation relative to the declared candidate relation $\mathcal{B}$. Requested-subject fidelity keeps z a function of $\mathcal{C}^\star$ alone, so satisfying both properties requires the separate channel or an equivalent relation label.

Proposition 1 (Licensed-set downward closure)**.** *If $c_1 \preceq c_2$ and $c_2 \in \mathcal{L}_\mathcal{K}(E)$, then $c_1 \in \mathcal{L}_\mathcal{K}(E)$.*

Proof. Downward closure of $\mathcal{A}(s)$ gives $c_1 \in \mathcal{A}(s)$. Obligation inheritance gives $\mathcal{O}(c_1) \subseteq \mathcal{O}(c_2)$, so every obligation required by c_1 passes under Π. □

Proposition 2 (Maximal coverage and status completeness)**.** *For a finite contract, $\mathcal{C}^\star$ and $\mathcal{R}^\star$ are antichains. Every licensed claim in $\downarrow c_{\mathrm{req}} \cap \mathcal{C}_{\sigma_{\mathrm{req}}}$ lies below an element of $\mathcal{C}^\star$; $\mathcal{C}^\star = \varnothing$ exactly when this requested-subject down-set contains no licensed claim. If c_{req} is licensed, then $\mathcal{C}^\star = \{c_{\mathrm{req}}\}$. The three values of z are mutually exclusive and exhaustive.*

Proof. Maximal elements form an antichain. Starting from any element of a finite candidate set and repeatedly moving to a strictly stronger candidate terminates at a maximal element. The remaining statements follow from the definition of $\mathcal{C}^\star$ and the three status cases. □

Proposition 3 (Licensing-favorable evidence refinement)**.** *For the same requested claim, let E and E' have the same source and satisfy $\bar{o}_\Pi(E) \leq \bar{o}_\Pi(E')$ for every $o \in \Omega$. Then $\mathcal{L}_\mathcal{K}(E) \subseteq \mathcal{L}_\mathcal{K}(E')$. For every $c \in \mathcal{C}^\star(E)$, some $c' \in \mathcal{C}^\star(E')$ satisfies $c \preceq c'$.*

Proof. Every obligation that passes for E also passes for E', which gives licensed-set inclusion. Finiteness extends each previously maximal requested-branch claim upward to a maximal claim in the refined licensed set. □

For the routing learned-primary branch $C0 \prec \cdots \prec Ck$, $\mathcal{C}^\star$ is empty or a singleton containing the highest licensed level; the empty case is represented by $L^\star = \bot$.

C.6 CLAIM-CARD SCHEMA

Every Claim Card contains a stable identifier, source-row pointer, score-only interpretation, claim subject, one strongest requested claim, required obligations, status, typed failures, key evidence, maximal supported claims, the retained set $\mathcal{R}^\star$, and an adjudication-record anchor. A collection of requested claims produces one execution record per request. Routing cards also expose the single highest licensed C-level produced by the evaluated chain.

Listing 2: Representative Claim Card: DataSP bridge transfer.

```
claim_id: C-04
source_row: DataSP multiseed bridge
score_only_interpretation: public transfer establishes preference-aware
    multigraph routing
claim_subject: learned routing component
requested_claim: C4
required_gates: [evidence_scope, hypervolume]
status: narrowed
failed_gates: [evidence_scope, hypervolume]
failure_type: source_mismatch_and_missing_evidence
key_evidence:
  learned_regret: 0.033519
  static_regret: 0.215459
  learned_hv: 0.303139
  static_hv: 0.329033
maximal_supported_claims: [C0]
retained_credit: []
highest_licensed_routing_level: C0
```

C.7 WHY MISSING TELEMETRY NARROWS CLAIMS

The primary policy Π_{cons} maps missing and unresolved required telemetry to zero, so the output records the scope licensed by the submitted evidence package. Intermediate and permissive policies alter selected non-core obligation resolvers while preserving source-scope, feasibility, fallback-attribution, multigraph-identity, and artifact-closure requirements. Section G.6 reports the resulting sensitivity analysis.

D VALIDATION DETAILS

This appendix expands the validation design summarized in Section 5. The layers separate conditional contract application, independent object construction, comparative reviewer utility, same-contract consequence, rule sensitivity, source-scope transfer, and artifact provenance.

Table 9: Validation layers and the distinct evidential role of each layer.

Layer	What it tests	Main anti-failure role
Fixed packets	Whether non-author experts apply the supplied contract consistently.	Conditional applicability of supplied objects.
Object construction	Whether non-authors reconstruct inputs from prespecified sections.	Upstream author dependence.
Reviewer comparison	Whether the method changes agreement, overpromotion, rationales, and time.	Practical benefit and conservative error.
Same-contract suite	Whether target-claim eligibility changes the selected row.	Consequence under shared reporting.
Scope and rule stress	Whether source credit and distinct gates behave as intended.	Transfer discipline and functional distinctness.
Versioned artifact	Whether core and extension-stage evidence can be traced and reanalyzed.	Chronological and computational closure.

D.1 REDUCED ADJUDICATORS AND ABLATIONS

We use contrastive adjudicators as diagnostic probes defined independently of the CLAIMSCOPE-120 labels. Each probe isolates information lost when the full claim-scope rule is simplified.

Table 10: Reduced adjudicators and rule ablations. The comparison separates three objects that are easy to conflate: reduced adjudicators, rule-necessity ablations, and the external human reference.

Probe	What is removed	Failure mode exposed
Regret-only ranking	Source scope, inherited gates, and record closure.	Selects Fallback-Heavy Safe in ROUTESCOPE-LITE because it has the lowest scalar regret; 86.1% fallback assigns the row to rescue-dominated hybrid evidence.
Missing-field checklist	Cumulative claim semantics and narrowing.	Can detect absent fields but cannot decide whether a row should be UNSUPPORTED, NARROWED, or credited at source-native scope.
Source-scope-only rule	Row-level telemetry gates.	Enforces bridge/native/out-of-contract transfer boundaries; fallback attribution, graph identity, dynamic behavior, and latency remain outside this reduced rule.
Leave-one-gate-out rules	One inherited gate at a time.	Each omitted gate enables a distinct false-support mode: fallback-heavy rescue, multigraph collapse, coverage loss, static equivalence, hidden latency, source upgrade, or unlocked artifacts.
Human-majority reference	Deterministic execution and manifest linkage.	Provides external claim-scope evidence, but is not a rerunnable row-level adjudicator.

The reduced rules isolate distinct parts of the licensing computation. Source-scope filtering enforces transfer boundaries for public shortest-path and native ML4CO evidence. Same-contract attribution and dynamic equivalence consume additional telemetry obligations. Missing-field prompts expose the evidence state, while the declared claim order and maximal-selection rule map that state to positive claim scope. The full rule composes source admissibility, inherited gates, attribution, and record closure.

D.2 CLOSED INTERNAL AUDIT PROTOCOL

The closed internal audit begins with a locked 22-row claim-relevant ledger. A row enters the seven-row headline registry if it satisfies at least one predeclared condition: it would ordinarily be read as a positive method result under scalar regret alone; it would naturally be interpreted as evidence for a stronger routing claim under prior benchmark practice; or omitting it would materially change the advertised scientific narrative. Rows outside the registry remain in the ledger as positive controls, stress evidence, diagnostics, efficiency upper bounds, or scoped bridge rows. The registry builder emits an exclusion reason for every non-registry claim-relevant row. This finite registry demonstrates the adjudication workflow on selected routing claims; literature-wide prevalence requires probability sampling.

The remaining validation details are distributed across the following sections: CLAIMSCOPE-120 sampling, construction, comparison, and annotation are described in Appendix E; ROUTESCOPE-LITE construction and uncertainty reporting are described in Appendix F; public source-scope rows, gate ablations, missing-telemetry policies, and additional result tables appear in Appendix G; artifact checks and failure semantics appear in Appendix I.

E CLAIMSCOPE-120 PROTOCOL

This appendix records the external-corpus protocol used to test whether SCOPE-ROUTING aligns with independent claim-scope judgments. The corpus provides a contract-relative operational reference for the specified evidence packets.

E.1 SAMPLING PROTOCOL

The CLAIMSCOPE-120 sampling protocol was fixed before adjudication. Papers were stratified into six source families: learned routing, decision-focused learning, ML4CO benchmarks, multiobjective optimization, evaluation-science frameworks, and bridge or out-of-contract tasks. Within each source family, claims were sampled from sections that commonly carry scientific interpretation: titles, abstracts, contribution lists, main results, and limitations. To avoid dominance by any single paper, each paper contributed at most five claims and three claims on average.

Claims were excluded when they were purely theoretical theorem statements, generic background claims, implementation descriptions without evaluation meaning, or claims whose evidence source could not be located in the public paper. The artifact records the exclusion reason for every screened candidate not included in the corpus.

E.2 CORPUS COMPOSITION AND AUDIT CHAIN

Table 11 reports the corpus composition. Table 12 gives the reviewer-facing audit chain. Each included claim has a stable identifier linking the candidate record, evidence excerpt, four raw labels, majority label, SCOPE-ROUTING label, missing-gate lists, and disagreement code. This trace supports reproducibility of the fixed-packet analysis; the separate construction audit below addresses upstream selection and encoding dependence.

Table 11: CLAIMSCOPE-120 corpus composition.

Stratum	Count	Sampling rule
Learned routing and neural routing	24	Public routing papers with empirical route-quality claims.
Decision-focused and predict-then-optimize learning	20	Claims about downstream decision quality, regret, or solver-aware training.
ML4CO benchmarks and native solver suites	22	Claims about benchmark coverage, solver-interface evidence, or native metrics.
Multiobjective optimization and preference learning	18	Claims involving hypervolume, coverage, Pareto quality, or preference conditioning.
Evaluation science, documentation, and stress testing	24	Claims about benchmark validity, intended use, stress tests, or evaluation frameworks.
External bridge tasks and out-of-contract comparators	12	Claims where task scope differs from SCOPE routing but evidence is relevant to source-scope behavior.
Total	120	40 papers; three claims per paper on average, capped to avoid domination by any single source.

Table 12: Reviewer-facing audit chain for CLAIMSCOPE-120.

Stage	Object or count	Reviewer check in the artifact
Candidate collection	Candidate claim records	Candidate IDs, source-family tags, claim location, inclusion/exclusion decision, and exclusion reason.
Corpus sample	120 claims from 40 papers	Stable Claim Cards with anonymized source identifiers and evidence excerpts.
Human annotation	Four blinded human labels per claim	Raw annotation matrix, missing-gate selections, anonymized annotator metadata, blinding/COI fields, and validation report.
Consensus construction	Majority label with tie rule	Scripted majority vote and tie-resolution log.
SCOPE-ROUTING adjudication	Executable label and missing gates	Adjudicator output linked to evidence source, telemetry/prose evidence, and artifact fields.
Disagreement audit	19 exact disagreements	Disagreement taxonomy and representative anonymized examples.

E.3 HUMAN-ANNOTATION BLINDING AND TASK

Each annotator saw the same restricted evidence packet: claim text, claim location, a short evidence excerpt, anonymized source-family label, and the annotation guide. Annotators did not see SCOPE-ROUTING outputs, internal registry labels, author-preferred labels, or other annotators' labels. They labeled the highest level licensed by the supplied evidence, selected missing gates from a fixed checklist, and marked disagreement causes when wording was ambiguous. When a claim was outside the routing contract, annotators could select OUT-OF-CONTRACT; when it was too vague to evaluate, they could select insufficient.

The artifact stores raw labels, anonymized annotator metadata, blinding status, conflict-of-interest screening fields, schema-validation status, consent and compensation materials, and ethics-review records. The expert-annotation protocol was submitted for institutional human-subjects determination and ethics review and was determined to be exempt. The annotation task involved adult domain experts making professional judgments about public scientific claims; it did not collect sensitive personal information, use deception, or expose participants to more than minimal risk. Annotators

were informed of the study purpose, data use, anonymization policy, and withdrawal option, and compensation was set to satisfy or exceed the applicable local minimum-wage standard for the estimated annotation time. To preserve double-blind review, institution-identifying protocol numbers are omitted from the anonymous manuscript and artifact; the camera-ready version will de-anonymize the review identifier if the paper is accepted. Personal identifiers are not included in the public package.

E.4 INDEPENDENT CONSTRUCTION AUDIT

The independent construction audit evaluates upstream object selection and encoding. We prespecified 24 claims and the paper sections available for encoding: abstract, method, experiment, results, and limitations. Two non-author coders independently reconstructed the claim span, evidence excerpt, comparator, source family, requested claim family, intended scope, required and missing gates, artifact availability, and final status while blind to author encodings and SCOPE-ROUTING outputs.

Table 13: Section-conditioned claim-evidence-object construction audit.

Field or outcome	Agreement	Interpretation
Comparator identity	23/24	Comparator extraction is highly reproducible.
Source family	22/24	Residual ambiguity concerns solver-interface credit.
Requested claim family	21/24	Dynamic wording creates the main family ambiguity.
Intended scope	weighted $\kappa = 0.78$	High but non-perfect ordinal agreement.
Required/missing gates	mean Jaccard $= 0.82$	Most evidentiary obligations are reconstructed similarly.
Final status	20/24 exact; 24/24 within one	Remaining differences are adjacent.

The predeclared resolution procedure changes four fixed-packet levels, all adjacently, and preserves the main conclusions. Remaining disagreements concern dynamic wording, latency, solver-interface source credit, and whether narrative script descriptions constitute a locked evidence record. The result measures section-conditioned construction reliability over the supplied paper sections.

E.5 HELD-OUT REVIEWER COMPARISON

The practical comparison uses 40 claims from 20 public papers that did not contribute to the ladder, gates, examples, or CLAIMSCOPE-120. Eight non-author experts evaluated claims under unstructured careful review, a structured evidence checklist, and SCOPE-ROUTING. Assignment followed a balanced incomplete-block design: every claim received two independent judgments per condition, no reviewer saw the same claim under multiple conditions, and condition instructions were fixed before analysis. A separate panel of three senior experts wrote the highest claim they considered licensed in free form while blind to the ladder and condition outputs; the panel resolved 36/40 claims and retained four unresolved cases in the artifact.

All conditions received the same prespecified abstract, method, experiment, results, and limitations sections. The checklist condition combined diagnostic prompts for comparator, feasibility, multigraph identity, preference coverage, fallback/attribution, dynamics, latency, source scope, artifact closure, and wording ambiguity with reviewer-authored final claims. The SCOPE-ROUTING condition supplied the same evidence fields together with the C0–C6 learned-primary inheritance relation and maximal-selection rule. The senior panel authored free-form references while blind to the branch and condition outputs. The study team subsequently mapped panel consensus statements through the prespecified contract-resolution log, producing an ontology-aligned expert-resolved operational reference shared by all three conditions.

The release retains all 240 judgment-level records, resolved condition summaries, free-form references, and the resolution log. A prespecified pair-resolution protocol produces the claim-level condition summary while retaining raw disagreements. Upward and downward diagnostics are separately aggregated from the paired judgments and can overlap with each other or with the resolved summary. A complete workflow output records the requested claim, source, comparator, observed evidence, failed obligations, and retained claim. Paper-clustered paired resampling gives an agreement difference of 19.4 points [5.6, 33.3] relative to careful review and 11.1 points $[-2.8, 25.0]$ relative to the checklist; the checklist interval spans zero. Assignments prevent same-claim repetition, while incomplete cross-claim order counterbalancing leaves possible cross-task learning.

For a descriptive cost analysis, let U_m and D_m be the separately recorded upward and downward events and define $B_r(m) = (rU_m + D_m)/36$. Because events may overlap, B_r is an event-burden index rather than a mutually exclusive final-decision loss. At the observed counts, SCOPE-ROUTING has lower event burden than careful review for $r > 1/2$ and lower burden than the checklist for $r > 1/3$. These point-estimate crossovers summarize the observed sample; cluster uncertainty and population-level cost preferences remain outside this calculation.

The extension artifact records instructions, assignments, anonymized raw judgments, condition summaries, reference statements, timing fields, conflict-of-interest checks, compensation records, applicable ethics documentation, encodings, and analysis scripts. These study records are anchored to the extension snapshot and reported with that provenance.

E.6 RELIABILITY CALCULATIONS

We treat C0–C6 as ordinal labels and treat OUT-OF-CONTRACT and insufficient as separate non-ordinal labels for exact agreement. Weighted Cohen's κ is computed pairwise on C0–C6 labels using quadratic weights and reported as the mean across annotator pairs. Claims marked OUT-OF-CONTRACT or insufficient by either annotator are included in exact agreement but excluded from ordinal κ. Krippendorff's α uses an ordinal distance over C0–C6 and a maximal distance to non-ordinal labels.

Agreement between SCOPE-ROUTING and humans is computed against a majority vote. The primary rule uses the lower median admissible ordinal level and the conservative source-scope label when a tied label indicates contract mismatch. Sensitivity analyses additionally use an upper median, a source-neutral resolution, and exclusion of tied cases. Across complete-corpus policies, exact agreement ranges from 0.833 to 0.850 and within-one agreement remains at least 0.950, showing stability across tie rules. Two exact disagreements lie at C3/C4 and two at C4/C5; the former concern coverage or fallback attribution, and the latter route-change, drift, or latency evidence.

For claim-level proportions we report Wilson intervals for comparability with the original analysis and paper-clustered bootstrap intervals in the revision artifact to account for multiple claims from the same paper. These intervals quantify finite-sample uncertainty. The majority label and senior-panel reference serve as contract-relative operational references for the specified evidence packets.

E.7 REPRESENTATIVE EXTERNAL-CORPUS DISAGREEMENTS

Table 14 shows that representative disagreements concentrate in interpretable boundary cases: source scope, deployment wording, artifact availability, preference coverage, and latency or dynamic evidence.

Table 14: Representative external-corpus disagreement cases. Source titles are anonymized in the submission version; artifact cards contain stable claim identifiers and source-family labels.

ID	Claim family	Human majority	SCOPE-ROUTING	Reason for disagreement
ECC-012	Shortest-path bridge	C1	C0	Humans credited feasibility from the shortest-path task; SCOPE-ROUTING kept source scope at C0 bridge because multigraph identity and fallback telemetry were absent.
ECC-019	Learned routing	C5	C4	Claim used dynamic language, but route-change and drift-stress telemetry were incomplete.
ECC-027	Multiobjective method	C3	C2	Strong scalarized results were reported, but hypervolume or preference coverage was missing.
ECC-038	Benchmark paper	C6	C5	Humans interpreted broad realistic coverage as deployment-style evidence; SCOPE-ROUTING required non-author usability and deployment telemetry.
ECC-044	Native ML4CO suite	C3	OUT-OF-CONTRACT	Native task instantiated multiobjective metrics but not source-destination multigraph routing.
ECC-061	Decision-focused learning	C0	C1	Human majority credited only scalar scoring; SCOPE-ROUTING found explicit feasible-decoding evidence in the excerpt.
ECC-073	Evaluation framework	C6	C6	Agreement case: framework-level evaluation guidance, not policy promotion.
ECC-084	Online adapter claim	C5	C4	Latency and route-change evidence were insufficient for dynamic robustness.
ECC-099	Routing heuristic	C2	C1	Edge identity was not preserved under a simplified graph representation.
ECC-117	Artifact claim	C4	insufficient	The paper described scripts but did not expose locked manifest or schema-valid telemetry.

E.8 ANNOTATION GUIDE

Task instructions. The human-annotation protocol used the following instruction text.

> You will label scientific claims from public ML, routing, ML4CO, decision-focused learning, multiobjective optimization, and evaluation papers. Focus on the strongest claim level supported by the supplied evidence excerpt. When evidence is missing, mark the missing gates. Assign OUT-OF-CONTRACT to claims outside the routing contract and insufficient to claims whose wording prevents evaluation.

Label definitions. The guide used the short definitions in Table 1, plus examples. It states that C0 is a valid positive claim when subject-bound evidence supports scalar learned-primary scoring, and that C6 is reserved for deployment-style claims with representative workloads, mechanism diagnosis, intended-use limits, and independent reproduction or usability evidence. A locked SCOPE-ROUTING adjudication record is an adjudication-validity requirement at every level and is not itself scientific evidence for C6.

Missing-gate checklist. Annotators selected all missing gates that applied:

- comparator or oracle not declared;

- feasibility conditional on nonempty feasible sets or invalid-route counts missing;
- multigraph edge identity missing;
- preference sampler or coverage missing;
- claim subject, fallback or repair taxonomy, or attribution evidence missing;
- route-change or drift evidence missing;
- latency or systems telemetry missing;
- source scope mismatch;
- artifact, manifest, or reproduction link missing;
- claim wording too ambiguous.

F ROUTESCOPE-LITE DETAILS

This appendix expands the same-contract stress suite used in Section 6. ROUTESCOPE-LITE tests how evidence-licensed evaluation changes conclusions under one routing contract. Its semisynthetic graph families and designed policy boundary cases provide mechanistic evidence; deployment evaluation requires live workloads and operational use records.

F.1 GRAPH AND REQUEST FAMILIES

ROUTESCOPE-LITE contains 1,500 requests across three semisynthetic domain families: urban-style multigraphs, logistics-style constrained routing graphs, and communication-style latency-reliability graphs. Each domain family contains five graph-family slices, for 15 slices total. Requests vary source-destination pair, preference vector, context, hard constraints, drift severity, and feasible-set size. Each slice is evaluated over 10 seeds. The request generator records oracle objective, feasible-set nonemptiness, preference vector, hard-constraint mask, and route identity.

F.2 POLICY FAMILIES

Table 15 lists the 15 policy families used in ROUTESCOPE-LITE. The inventory spans boundary cases for attribution, identity, coverage, dynamics, and latency. The adapter interface is shared across families: score edges, retain candidate edges, decode a path, optionally fallback or repair, and emit telemetry.

Table 15: Policy-family inventory for ROUTESCOPE-LITE.

Policy family	Type	Evidentiary role
Dynamic exact oracle	upper bound	Provides oracle quality and feasibility reference.
Static exact anchor	comparator	Strong nonlearned comparator for regret and route-change checks.
Weighted shortest path	heuristic	Classical baseline with explicit weights.
k-shortest-path heuristic	heuristic	Tests nonlearned multigraph candidate generation.
Simple-graph collapse	negative control	Tests C2 multigraph identity gate.
Learned full-exact scorer	learned scoring	Tests feasibility-audited controlled learned-scoring claims.
Learned top-k pruner	learned pruning	Tests retained-edge failure and fallback masking.
Fallback-heavy safe policy	safety control	Tests whether low regret from fallback is blocked.
Online adapter	adaptive policy	Tests route-change, latency, and nonzero-effect gates.
RL-style route policy	learned policy	Tests broader policy-family compatibility.

Policy family	Type	Evidentiary role
Preference hypernetwork	learned preference policy	Tests C3 coverage and held-out preference shift.
DFL bridge adapter	external bridge	Tests source-scope limits under same reporting surface.
Random scorer	negative control	Tests plumbing and failure semantics.
Adaptive pruning repair	repair policy	Tests whether targeted fixes improve supported claim level.
Shock-safe online repair	repair policy	Tests C5 dynamic robustness without C6 promotion.

F.3 Uncertainty analysis

Aggregate rows report seed-level confidence intervals, request-paired bootstrap intervals for regret differences, fallback-rate intervals, aggregate and conditional feasibility counts, hypervolume intervals, route-change win/tie/loss, and p50/p95/p99 latency on the declared hardware. The 1,500 requests are the union across 15 graph slices and 10 seeds; the artifact records the per-slice and per-seed counts and retains 1,500 as the independent-request count. The main text reports point estimates for readability, while the manifest indexes the complete uncertainty table and hardware record.

F.4 Selected policy rows and uncertainty tables

Tables 16–18 give the claim-relevant rows and uncertainty audit used for the conclusion-change claim. Table 15 enumerates all 15 families; the manifest-linked CSV contains all 15 quantitative rows, including Weighted Shortest Path, Online Adapter, and DFL Bridge Adapter, which are omitted from the printed table only for space.

Table 16: Selected ROUTESCOPE-LITE rows. "Feas." is the reported aggregate rate; C1+ uses exact conditional invalid-route counts.

Policy family	Regret	Fallback	Feas.	HV	Licensed interpretation
Dynamic exact oracle	0.000000	0.000	1.000	0.892	Upper bound; not ranked as learned policy.
Fallback-Heavy Safe	0.011091	0.861	0.997	0.669	Standalone C5 not licensed; hybrid safety-control credit retained in $\mathcal{R}^\star$.
GNN Pruned Top-$k = 1$	0.012508	0.505	0.955	0.439	C0; fallback and coverage gates fail.
Static exact anchor	0.014874	0.000	1.000	0.716	Comparator anchor, not learned routing.
Shock-Safe Online Repair	0.016189	0.012	1.000	0.819	C5-eligible under the declared claim-conditioned rule.
Adaptive Pruning Repair	0.016761	0.016	1.000	0.799	C4; repair improves pruning claim level.
GNN Full Exact	0.018025	0.000	1.000	0.789	C4 feasibility-audited learned scoring.
Preference Hypernetwork	0.020865	0.003	0.998	0.849	C4; coverage-preserving preference-conditioned controlled policy.
RL Route Policy	0.023410	0.041	0.991	0.764	C3; dynamic evidence incomplete.
k-Shortest-Path Heuristic	0.027804	0.000	1.000	0.731	C2; nonlearned multigraph heuristic baseline.
Simple-Graph Collapse	n/a	0.000	0.884	n/a	Rejected for C2+ multigraph claims.
Random Scorer	0.064220	0.183	0.912	0.352	Negative-control plumbing check.

Table 17: Predeclared claim-conditioned selection rules for ROUTESCOPE-LITE; each rule ranks only its eligible policy rows.

Selection rule	Selected row	Reason
Regret only	Fallback-Heavy Safe	Lowest aggregate regret, without an attribution eligibility requirement.
Target C5, then regret	Shock-Safe Online Repair	Passes inherited C5 gates; eligible rows are then ordered by regret.
Target C5, then normalized cost	Shock-Safe Online Repair	Remains selected after the declared latency and cost normalization.
Coverage-first diagnostic	Preference Hypernetwork	Highest learned non-oracle HV, but incomplete C5 dynamics.

Table 18: ROUTESCOPE-LITE uncertainty audit. The conclusion change follows from a declared target claim, eligibility evidence, and an explicit within-eligible quality rule.

Quantity used in claim	Point estimate in main table	Required uncertainty check
Regret order between FallbackHeavySafe and ShockSafeOnlineRepair	0.011091 vs. 0.016189	Request-paired bootstrap interval for regret difference; used to characterize the regret-only ranking before claim-conditioned eligibility.
Fallback attribution	0.861 vs. 0.012	Seed- and request-level fallback intervals plus primary-only/end-to-end decomposition; narrowing depends on the aggregate gain remaining rescue-dominated.
Feasibility boundary	0.997 vs. 1.000	Aggregate interval and exact invalid-route counts conditional on nonempty feasible sets; the exact conditional counts determine C1.
Coverage comparison	0.669 vs. 0.819 HV	Hypervolume interval; C3+ support uses nondegraded preference-space evidence alongside scalar regret.
Route-change and dynamic evidence	artifact-reported W/T/L	Route-change win/tie/loss and non-static-equivalence checks; C5 support uses observed dynamic behavior beyond inherited static-anchor routes.
Latency and cost-conditioned selection	artifact-reported p50/p95/p99	Latency intervals, hardware record, normalization, weights, and tie handling; operational selection depends on these records.

G Additional Results and Robustness

This appendix expands the headline results in Section 6. The main text reports the takeaways; here we provide the full agreement tables, public firewall, internal registry summaries, positive calibration rows, deployment boundary, gate ablations, missing-telemetry policies, and claim-card examples.

G.1 Representative held-out interpretation changes

Table 19 preserves the three representative interpretation changes summarized in Section 6. These examples show same-subject narrowing and source- or task-scope boundaries independently of the fallback-heavy cross-subject witness.

Table 19: Representative interpretation changes in held-out public-paper claims.

Requested claim	Observed evidence gap	Evidence-supported effect
Identity-preserving multi-graph routing	Evaluation collapses parallel edges to a simple graph	Multigraph claim blocked; lower structural or scalar credit may remain.
Preference-conditioned routing	Scalar quality reported without preference-space coverage	$C3$ blocked; scalar and inherited lower-scope credit retained.
Dynamic robust routing	No route-change or drift evidence	$C5$ blocked; supported static or attribution-audited credit retained.

G.2 External claim-corpus results

Table 20 reports external agreement statistics. Table 21 gives the collapsed confusion matrix. Table 22 summarizes the 19 exact disagreements.

Table 20: Conditional fixed-packet agreement using blinded non-author annotation. Human majority provides the operational reference under the specified guide and evidence packets.

Quantity	Value	Interpretation
Claims / papers / domains	120 / 40 / 6	Public corpus outside the internal SCOPE ledger.
Annotators	4 blinded non-author experts	ML4CO/routing, evaluation/benchmarking, OR/optimization, and ML systems/reproducibility perspectives.
Mean pairwise exact agreement	0.761	Experts often agree exactly on non-trivial claim-scope labels.
Mean pairwise within-one-level agreement	0.931	Most disagreements occur at adjacent scope levels.
Weighted Cohen's κ (mean pairwise)	0.703	Substantial ordinal agreement.
Krippendorff's α (ordinal)	0.686	Conservative multi-annotator reliability estimate.
SCOPE-ROUTING exact vs majority	101/120 = 0.842; 95% Wilson CI [0.766, 0.896]	Agreement under fixed packets and the specified contract.
SCOPE-ROUTING within-one vs majority	115/120 = 0.958; 95% Wilson CI [0.906, 0.982]	Most disagreements are adjacent.
Tie-policy sensitivity	exact 0.833–0.850; within-one ≥ 0.950	Conservative, upper-median, source-neutral, and tied-case-excluded policies preserve the conclusion.
SCOPE-ROUTING exact vs each annotator, mean	0.803	Comparable to inter-annotator exact agreement.
Human-annotation audit fields	Complete	Raw labels, anonymized annotator metadata, blinding/COI fields, consent and compensation materials, and exempt ethics-review records validate in the artifact; personal identifiers are excluded from the public package.

Table 21: Collapsed confusion matrix for SCOPE-ROUTING vs majority human label on CLAIMSCOPE-120. Rows are majority labels; columns are SCOPE-ROUTING labels.

Majority group	C0/bridge	C1–C2	C3–C4	C5–C6	Off/insufficient
C0/bridge	25	2	0	0	1
C1–C2	3	21	2	0	1
C3–C4	1	2	32	2	0
C5–C6	0	0	3	15	2
Off/insufficient	0	0	0	0	8

Table 22: Disagreement taxonomy for the 19 exact disagreements between SCOPE-ROUTING and the majority human label.

Disagreement type	Count	Typical pattern
Ambiguous deployment wording	6	The paper uses realistic or practical language without C6-scale evidence.
Source-scope strictness	5	Humans credit bridge or native-task evidence one level higher than the source-scope rule allows.
Artifact availability	4	Humans infer reproducibility from prose, while SCOPE-ROUTING requires manifest-linked objects.
Preference coverage boundary	2	Scalar preference results are strong but coverage evidence is partial.
Latency/online-effect boundary	2	Dynamic or online language appears, but route-change or latency telemetry is incomplete.

G.3 PUBLIC SCOPE FIREWALL

The Scope Firewall audits real public papers and benchmark artifacts whose task contracts differ from dynamic preference-conditioned multigraph routing. These cases are intentionally not treated as same-contract SCOPE routing baselines. Each audit separates native/source-supported credit from SCOPE-transfer scope.

Table 23: Public Scope Firewall. Native/source-supported credit is preserved, while SCOPE-transfer scope remains bounded by the dynamic preference-conditioned multigraph routing contract.

Public source	Native/source-supported credit	SCOPE-transfer scope	Boundary reason
PyEPO Tang & Khalil (2022)	Predict-then-optimize library evidence across shortest path, knapsack, TSP, and LP/IP tasks.	Bridge/native decision-quality evidence only.	Does not instantiate the SCOPE dynamic multigraph route object or its telemetry contract.
DATASP Lahoud et al. (2024)	Contextual cost learning and differentiable all-to-all shortest-path/path-prediction evidence.	Shortest-path bridge evidence with coverage caveat.	Parallel-edge identity, SCOPE fallback taxonomy, preference-shift coverage, and dynamic route-change telemetry are absent.
FRONTIERCO Feng et al. (2026)	Native large-scale CO solver benchmark/readiness evidence; local TSP diagnostics are executable.	Native diagnostic/readiness, not same-contract SCOPE routing.	No SCOPE source-destination route-producing policy and no SCOPE route telemetry.
CO-Bench Sun et al. (2026)	Native LLM-agent algorithm-search benchmark; RCSP native solver-interface audit solves or certifies local source-scope review instances and derived records.	Solver-interface/source-scope evidence; no learned multigraph-routing escalation.	Evaluated object is algorithm search or native RCSP solving, not a SCOPE route-producing policy.
PredOpt-SP	Reproduced shortest-path decision-focused bridge comparators.	C0 shortest-path bridge audit.	Not strict same-walltime evidence and not dynamic preference-conditioned multigraph routing.
PredOpt-Knapsack	Native constrained-optimization closure over local selector and official comparator rows.	Off-thesis positive evidence, not routing evidence.	Knapsack is not a source-destination route contract.
Artifact trace	`artifacts/public_end_to_end_audit_case_20260503/` and `artifacts/full_external_benchmark_audit_20260504/`.	Claim Cards, gate traces, executable source-scope rows, and manifest records.	Rebuilt by the reviewer artifact.

Table 24: Executable external benchmark audit. These rows support native, bridge, out-of-contract, or diagnostic claims; they are not promoted into same-contract SCOPE routing unless the SCOPE route contract is instantiated.

Evidence block	Coverage	Claim boundary
PredOpt-SP	Latest local official-code extension plus preference-function bridge rows; reproduced SPO+ bridge remains stronger than the local selector row on the audited bridge slice.	Shortest-path bridge audit; blocks all-baseline or state-of-the-art routing claims.
PredOpt-Knapsack	Local selector rows and official comparator rows across method files; the selector-family mean is better than the best comparator-family mean on the native constrained-optimization audit.	Full native PredOpt knapsack closure, but out-of-contract for routing.
DATASP	Three-seed SCOPE bridge over fixed synthetic episodes plus SPO+ bridge baseline.	C0 shortest-path bridge with coverage caveat, not native full DATASP routing validation.
FRONTIERCO TSP	Native TSP/LKH diagnostics and seed portfolio; no SCOPE-native tour-producing policy.	Native diagnostic/readiness evidence only under unmatched task contract.
CO-Bench RCSP bridge	Subset of raw RCSP instances converted through a single-resource SCOPE bridge; the subset is saturated by static/exact baselines.	RCSP bridge coverage audit; no learned-routing win claimed.
CO-Bench RCSP native audit	All local source-scope review instances are audited; each instance is solved optimally or certified infeasible under the native solver-interface check, without redistributing upstream raw assets whose reuse terms are unclear.	Native solver-interface/source-scope evidence; C2–C6 SCOPE routing promotion remains blocked.

G.4 INTERNAL REGISTRY AND POSITIVE CALIBRATION

The closed internal registry is a finite demonstration over locked claim-relevant rows. Table 25 reports the registry, and Table 26 shows that the same adjudicator preserves valid positive credit when gates pass. Population-rate estimation requires a probability sample.

Table 25: Closed internal registry used for the finite demonstration; population-rate estimation requires probability sampling.

Claim family	Regret-only interpretation	Status	Deciding evidence and supported scope
Aggressive pruning	Top-$k = 1$ appears attractive because it reaches regret 0.009552.	UNSUPPORTED	Fallback rises to 40.94% and hypervolume falls to 227.985; no positive pruning or operational routing claim.
Simple-graph collapse	Feasible-only regret could suggest that collapse is acceptable.	UNSUPPORTED	Feasibility falls to 90.6% on the controlled slice and 50.0% on the diagnostic slice; C2+ requires edge identity.
Realistic diagnostic learned win	Equality with a strong static anchor could be read as realistic stability.	UNSUPPORTED	Equality is mediated by 100% fallback and route-static behavior; supported claim is diagnostic failure analysis.
DATASP bridge	Scalar regret improves from 0.21546 to 0.03352.	NARROWED	Hypervolume decreases from 0.32903 to 0.30314 and the source is not multigraph; supported claim is learned-scoring bridge evidence with a coverage caveat.
FRONTIERCO diagnostics	External compatibility could be read as native routing success.	UNSUPPORTED	The requested routing branch is empty; $\mathcal{R}^\star$ retains compatibility and solver-interface efficiency diagnostics.
Public Scope Firewall	Real public benchmark evidence could be overread as same-contract routing validation.	NARROWED	Public rows retain native/source-supported credit while SCOPE-transfer is blocked when the dynamic multigraph route contract is absent.
Online adapter	Adapter-on appears harmless on ordinary streams.	UNSUPPORTED	Quality is tied while DATASP step latency increases by about 12.323 ms; no positive online-adaptation claim.
Preference conditioning	Preference input appears necessary by design.	UNSUPPORTED	The ordinary-split request is unsupported; $\mathcal{R}^\star$ retains held-out preference-shift robustness.

Table 26: Positive calibration on the controlled multigraph slice. The evaluator preserves valid positive credit while blocking unsupported escalation.

Row	Regret	Δ vs static	Feas.	Fallback	Adjudicated claim
Static exact anchor	0.014154	–	100%	0.000000	Static exact comparator.
Learned full-exact scorer	0.009878	-30.2%	100%	0.000000	SUPPORTED to C0–C4: feasibility-audited controlled learned scoring.
Top-$k = 4$ retained decode	0.009931	-29.9%	100%	0.000000	NARROWED to top-k-specific retained-decode evidence.
Collapsed exact control	n/a	–	90.6%	0.000000	UNSUPPORTED for C2+ multigraph claims.

G.5 DEPLOYMENT BOUNDARY

C6 requires deployed multi-domain workloads, live operational telemetry, non-author community submissions, and intended-use validation. The available evidence licenses controlled, scoped public, and same-contract dynamic claims through C5.

Table 27: Deployment boundary: available evidence licenses controlled claims through C5, while C6 requires deployment evidence.

Evidence block	Highest licensed level	Missing C6 evidence	Interpretation
Controlled internal audit	C4	Multi-domain realistic support, dynamic robustness, non-author usability.	Valid controlled learned-scoring evidence only.
ROUTESCOPE-LITE semisynthetic suite	C5	Deployed topology, live workload, non-author reproduction, intended-use validation.	Strong same-contract stress evidence, not deployment.
Diagnostic realistic slice	DIAGNOSTIC only	Sample size, fallback-free learned win, multi-domain realism.	Mechanism evidence for fallback/static failures.
External public rows	Scoped public credit	Same routing contract, telemetry completeness, route identity.	External scope firewall evidence.
Public Scope Firewall	Scoped source credit	Dynamic preference-conditioned multigraph routing contract and telemetry.	Real public sources receive native, bridge, solver-interface, or out-of-contract credit, not deployment or multigraph-routing escalation.
Artifact dry run	Record closure	Live external use, public non-author adapters, and deployment evidence.	Reviewer-runnable analysis, not community adoption.

G.6 GATE ABLATIONS AND MISSING-TELEMETRY SENSITIVITY

Table 28 reports leave-one-gate-out stress. Each removed gate enables a distinct false-support mode. Table 29 reports conservative, intermediate, and permissive missing-telemetry policies. The strongest boundary results depend on source scope, feasibility auditing, fallback attribution, multigraph identity, and record closure.

Table 28: Gate ablations. Each removed gate enables a distinct false-support mode.

Gate removed or stressed	Overcredit events	False-support mode blocked by full rule
Fallback	4	Fallback-heavy pruning, fallback-heavy same-contract winners, or diagnostic equality would be credited as learned routing.
Multigraph identity / feasibility	3	Simple-graph collapse would be credited as valid multigraph routing.
Preference coverage	3	Scalar regret would be credited as preference-conditioned coverage despite HV loss.
Route change	3	Static-equivalent or no-effect online rows would be credited as dynamic behavior.
Latency	2	Operational claims would ignore slower or impractical paths.
Source scope	5	Bridge, native, or out-of-contract evidence would be silently upgraded to full routing.
Record lock	12	Registry and external-corpus conclusions would no longer be bound to locked rows, Claim Cards, and manifest entries.

Table 29: Sensitivity policies for missing telemetry. The manuscript's main results use the conservative policy; the artifact can report intermediate and permissive variants as a robustness check.

Policy	Missing-evidence treatment	Boundary results expected to remain unchanged
Conservative SCOPE-ROUTING	Any inherited missing gate blocks support at the affected level; lower supported levels may remain.	Main paper rule.
Intermediate SCOPE-ROUTING	Record, source-scope, feasibility, fallback-attribution, and multigraph-identity requirements remain hard; latency and partial coverage gaps usually narrow.	Rescue-dominated rows still cannot become standalone learned-routing evidence; semisynthetic rows still cannot become C6.
Permissive SCOPE-ROUTING	Missing non-core telemetry narrows by one or more levels when source scope is compatible and no hard requirement fails.	Bridge/out-of-contract evidence still cannot become full multigraph routing; unlocked records still cannot support adjudicated claims.

G.7 INTERNAL CLAIM-CARD EXAMPLES

Each internal registry row has a machine-readable Claim Card. Listings 3–9 are human-readable versions. They illustrate the path from a score-only interpretation to routing status and retained source-native or diagnostic credit.

Listing 3: Claim card C-01: aggressive pruning headline row.

```
claim_id: C-01
source_row: controlled multigraph top-k=1
score_only_interpretation: aggressive pruning works as a positive routing
    claim
claim_subject: learned pruning component
requested_claim: C5
required_gates: [fallback, hypervolume, latency]
status: unsupported
failed_gates: [fallback, hypervolume, latency]
failure_type: observed_failure_and_attribution_gap
```

```
key_evidence:
  regret: 0.009552
  fallback: 0.409375
  hypervolume: 227.985
maximal_supported_claims: []
retained_credit: [negative_control_fallback_and_coverage]
highest_licensed_routing_level: none
```

Listing 4: Claim card C-02: simple-graph collapse.

```
claim_id: C-02
source_row: collapsed exact control
score_only_interpretation: simple-graph collapse is acceptable for
    multigraph routing
claim_subject: collapsed routing system
requested_claim: C2
required_gates: [multigraph_feasibility]
status: unsupported
failed_gates: [multigraph_feasibility]
failure_type: observed_structural_failure
key_evidence:
  controlled_feasibility: 0.906
  diagnostic_feasibility: 0.500
maximal_supported_claims: []
retained_credit: [action_space_collapse_diagnostic]
highest_licensed_routing_level: none
```

Listing 5: Claim card C-03: realistic diagnostic learned win.

```
claim_id: C-03
source_row: realistic diagnostic learned row
score_only_interpretation: realistic-looking equality demonstrates
    learned dynamic routing
claim_subject: learned routing component
requested_claim: C6
required_gates: [fallback, route_change, mechanism_clarity]
status: unsupported
failed_gates: [fallback_credit, route_change, mechanism_clarity]
failure_type: attribution_gap_and_observed_static_equivalence
key_evidence:
  regret: 0.000055
  fallback: 1.000000
  selected_not_static_before_scan_top_k: 0.000
maximal_supported_claims: []
retained_credit: [failure_mechanism_diagnostic]
highest_licensed_routing_level: none
```

Listing 6: Claim card C-04: DataSP bridge transfer.

```
claim_id: C-04
source_row: DataSP multiseed bridge
score_only_interpretation: public transfer establishes preference-aware
    multigraph routing
claim_subject: learned routing component
requested_claim: C4
required_gates: [evidence_scope, hypervolume]
status: narrowed
failed_gates: [evidence_scope, hypervolume]
failure_type: source_mismatch_and_missing_evidence
key_evidence:
  learned_regret: 0.033519
  static_regret: 0.215459
  learned_hv: 0.303139
  static_hv: 0.329033
```

```
maximal_supported_claims: [C0]
retained_credit: []
highest_licensed_routing_level: C0
```

Listing 7: Claim card C-05: FrontierCO compatibility.

```
claim_id: C-05
source_row: FrontierCO bridge/native diagnostics
score_only_interpretation: compatibility rows imply native external
    routing success
claim_subject: external compatibility evidence
requested_claim: C6
required_gates: [native_task_parity, runtime_scope]
status: unsupported
failed_gates: [native_task_parity, runtime_scope]
failure_type: source_mismatch
key_evidence:
  route_bridge_status: tied_quality_slower_runtime
  pi_file_reuse: same objective with runtime reduction only
  equal_budget_portfolio: upper-bound controller evidence with deployment
      gates absent
maximal_supported_claims: []
retained_credit: [compatibility_diagnostic, solver_interface_efficiency]
highest_licensed_routing_level: none
```

Listing 8: Claim card C-06: online adapter benefit.

```
claim_id: C-06
source_row: ordinary-stream online adapter evaluation
score_only_interpretation: online adapter improves or harmlessly
    preserves routing quality
claim_subject: online adaptation component
requested_claim: C5
required_gates: [nonzero_effect, route_change]
status: unsupported
failed_gates: [nonzero_effect, route_change]
failure_type: observed_no_effect
key_evidence:
  ordinary_stream_quality_delta: 0.0
  datasp_added_step_latency_ms: 12.323
maximal_supported_claims: []
retained_credit: [no_effect_diagnostic, latency_diagnostic]
highest_licensed_routing_level: none
```

Listing 9: Claim card C-07: ordinary preference-conditioning necessity.

```
claim_id: C-07
source_row: ordinary preference-conditioning ablation
score_only_interpretation: preference input is necessary by design on the
     ordinary split
claim_subject: preference-conditioned routing component
requested_claim: C3
required_gates: [held_out_preference_shift]
status: unsupported
failed_gates: [ordinary_split_necessity]
failure_type: evidence_slice_mismatch
key_evidence:
  ordinary_split_signal: weak_or_noisy
  held_out_shift_pref_full_exact_regret: 0.002379
  held_out_shift_pref_full_exact_hv: 114.061993
  held_out_shift_uniform_pref_regret: 0.029327
  held_out_shift_uniform_pref_hv: 59.817196
maximal_supported_claims: []
retained_credit: [C3_on_held_out_preference_shift]
highest_licensed_routing_level: none
```

H Extended Implications and Limitations

This appendix expands the implications and scope statements summarized in Section 7.

H.1 Implications for authors, reviewers, and benchmark builders

For authors, each benchmark row pairs its score with a requested claim, claim subject, source family, telemetry, comparator, and adjudication record. A fallback-heavy row can retain hybrid safety-control credit while standalone learned-routing attribution remains unsupported; a bridge task can support learned-scoring transfer at its source scope; and a feasibility-audited scorer can receive precisely scoped positive credit.

For reviewers, the central question is which claim the full evidence package licenses under a declared contract. ClaimScope-120 measures conditional agreement on fixed packets, the construction audit tests upstream encoding, and the held-out comparison measures practical adjudication trade-offs. Together they make the inputs, boundaries, and rationales of scientific judgment more observable.

For benchmark builders, source scope, telemetry, attribution, and record closure are part of the evidence needed to interpret a row. An adapter emits observable fields; a Claim Card declares the request and source; the evaluator returns the highest licensed routing level and the incomparable retained set $\mathcal{R}^\star$. Domain-specific gates must be revised when the task changes.

H.2 Broader impacts

The main positive impact of SCOPE-Routing is methodological: it encourages authors, reviewers, and benchmark builders to record claim scope explicitly alongside benchmark scores. By turning source scope, telemetry requirements, and artifact closure into auditable objects, the framework can reduce overclaiming, improve reproducibility, and make benchmark results easier to inspect and contest.

Potential negative effects also exist. A conservative adjudicator may under-credit methods whose evidence is incomplete while their underlying behavior exceeds the reported record, and mechanically reusing the same claim-support ladder outside dynamic preference-conditioned multigraph routing could lead to inappropriate downgrades or misleading conclusions. Claim-scope tools work as structured aids to scientific judgment; domain changes call for redesigned evidence requirements and a newly validated contract.

H.3 Expanded limitations

Routing-specific gates require adaptation. The claim-support ladder is designed for dynamic preference-conditioned multigraph routing. Some gates, such as multigraph identity and route-change win/tie/loss fields, are domain-specific. Appendix B illustrates a selection-task instantiation; empirical cross-domain validation remains a future study.

Contract-relative operational references. ClaimScope-120 uses completed blinded non-author annotations as an operational reference under the supplied evidence packets and guide. The corpus contains 120 claims from 40 papers, expertise is finite, and excerpts simplify full-paper context. The 24-object audit is section-conditioned, while the 40-claim comparison has four unresolved reference cases and limited power against the checklist. These studies establish finite-sample execution reliability and trade-offs for the declared contract.

Semisynthetic mechanistic evidence. RouteScope-Lite provides same-contract breadth across policies, graph families, requests, and seeds. C6 additionally requires deployed topology, live workloads, operational constraints, and independent use evidence; the present records therefore license levels through C5. This boundary illustrates how the evaluator keeps controlled evidence within its observed scope.

Prevalence requires probability sampling. The public source-scope rows and claim corpora are purposive samples. They expose recurring, identifiable failure modes and demonstrate controlled

conclusion changes. Estimating literature-wide frequency and real-world impact requires a probability sample and broader task coverage.

Source-scoped public audit. The executable audit adds PredOpt-family rows and a CO-Bench RCSP solver-interface check over local review instances and derived summaries. These tasks receive source-native credit, while same-contract learned multigraph-routing claims remain tied to the routing telemetry contract. Upstream assets with unclear reusable licenses remain outside the redistributed package.

Adoption evidence. The artifact exposes reviewer commands and local checks, and the held-out study measures paper-adjudication time. Community adoption additionally calls for public non-author adapter submissions, longitudinal use, and independent usability studies.

Reduced adjudicators are diagnostic probes. The reduced adjudicators and leave-one-gate-out variants expose what information is lost when the full rule is simplified. They provide constructed witnesses showing that regret-only ranking, missing-field checking, and source-scope filtering leave different claim-scope errors unresolved. The resulting design lesson is to represent score, source scope, telemetry, claim-order semantics, and artifact closure jointly, with a newly authored contract for each task family.

Adjudication can be conservative. SCOPE-ROUTING treats missing telemetry as missing evidence. This may under-credit papers whose methods satisfy a property that the report leaves undocumented. The sensitivity policies in Section G.6 make this choice explicit, and the resulting output describes the scope licensed by the available record.

I ARTIFACT AND REVIEWER PATH

The artifact is part of the evidence for SCOPE-ROUTING: claim licensing depends on locked summaries, schema-valid telemetry, Claim Cards, manifest entries, and versioned analysis scripts. We distinguish the May 5, 2026 core snapshot `core_snapshot_20260505` from analyses first recorded in the July 28, 2026 extension snapshot `extension_snapshot_20260728`. Later evidence is not retroactively attributed to the core snapshot.

Table 30: Artifact chronology and execution mode. "Fresh analysis" recomputes results from released raw records; "cached reanalysis" recomputes summaries from locked experimental outputs; "derived-only" preserves historical summaries without silently rerunning external or heavy experiments.

Evidence object	Core snapshot	Extension snapshot	Execution mode
Adjudicator, schema, Claim Cards, registry	yes	unchanged/core	Fresh validation and table/card generation.
CLAIMSCOPE-120 packets and four-expert labels	yes	extended reporting	Fresh reanalysis from released labels and tie logs.
ROUTESCOPE-LITE request and policy rows	yes	unchanged/core	Cached reanalysis from locked rows; not model retraining.
Public source-scope audits	yes	extended summaries	Fresh local checks or derived source-scoped summaries as documented.
40-claim reviewer comparison	no	added	Fresh analysis from anonymized raw judgments, assignments, and times.
24-object construction audit	no	added	Fresh analysis from independent encodings and resolution logs.
Heavy training and external solver sweeps	historical only	historical only	Derived-only summaries; not rerun by the default path.

I.1 ARTIFACT LAYOUT

Listing 10 gives the versioned artifact layout used throughout the paper. Directory and script names follow the SCOPE/ClaimScope/RouteScope terminology consistently.

Listing 10: Artifact directory layout.

```
scope-routing-artifact/
  README.md
  QUICKSTART.md
  run_all.sh
  Dockerfile
  environment.yml
  Makefile
  LICENSE
  LICENSE-CODE
  THIRD_PARTY_NOTICES.md
  DATA_CARD.md
  configs/
    promotion_rule.yaml
    evidence_map.json
    claim_registry.json
    comparators.yaml
  data/
    claimscope_120/
    heldout_reviewer_comparison/
    object_construction_audit/
    routescope_lite/
  src/
    adjudicator/
    exporters/
    adapters/
  scripts/
    validate_artifact.sh
    rebuild_claim_registry.py
    reproduce_main_tables.sh
    print_claim_cards.py
    run_smoke_test.sh
    check_adapter_compliance.py
    run_claimscope_120.py
    run_heldout_reviewer_comparison.py
    run_object_construction_audit.py
    build_source_scope_audit_20260504.py
    run_cobench_rcsp_full_native_audit_20260504.py
    run_routescope_lite_reanalysis.sh
  artifacts/
    locked_summaries/
    claim_cards/
    public_end_to_end_audit_case_20260503/
    extended_source_scope_audit_20260504/
    cobench_rcsp_full24_native_audit_20260504/
    reproduction_reports/
  metadata/
    artifact_manifest.json
    checksums.sha256
```

I.2 ANONYMOUS REPOSITORY

For reviewer convenience, the submission also provides an anonymous read-only repository mirror at `https://anonymous.4open.science/r/SCOPE-ROUTING-B22D`. The mirror exposes the versioned artifact structure referenced throughout this appendix and is intended as a navigation aid for code, configuration files, Claim Cards, anonymized study records, and analysis scripts. The manifest identifies which objects belong to `core_snapshot_20260505` and `extension_snapshot_20260728`.

I.3 CONSISTENCY INVARIANT

For every reported row, the identifier must match one locked summary, manifest entry, Claim Card, evidence-map scope, and snapshot designation. Every held-out or construction-audit statistic must trace to anonymized raw records, assignments, encodings, and one analysis output. Every claim-relevant row outside the closed registry must have one exclusion reason. Validation fails on missing objects, duplicate mappings, cross-snapshot attribution, or a numerical table that cannot be regenerated from its designated raw or cached input.

I.4 LICENSES AND TERMS OF USE

The submitted artifact separates newly authored SCOPE-ROUTING components from third-party assets used only for source-scope auditing. Newly authored adjudicator code, schemas, exporters, adapters, and reviewer scripts are released under an explicit software license in `LICENSE-CODE`; newly generated Claim Cards, manifests, documentation, and aggregate audit summaries are released under the artifact documentation/data license in `LICENSE`. Third-party assets are not relicensed by this submission. Their source, access date, version or commit when available, and license or terms-of-use status are recorded in `THIRD_PARTY_NOTICES.md` and in the manifest.

This is a *yes, with restrictions* licensing posture: creators and original sources are credited, license or terms-of-use status is documented per asset, and assets with unknown or absent reusable licenses are kept out of the redistributable and default-executable artifact. The public package therefore supports the paper's experiments through independently authored code, generated telemetry, Claim Cards, manifests, and aggregate audit summaries, not by repackaging third-party code or raw data whose reuse terms are unclear.

Table 31: License and terms-of-use status for external assets touched by the source-scope audit, as checked on the official project pages on May 4, 2026. "Not redistributed" means that the artifact does not vendor, mirror, or relicense the upstream code or raw data; only independently authored scripts, source citations, Claim Cards, and derived aggregate audit records are included.

Asset	Asset type	Source-provided status	Restricted use in this submission
PyEPO	Code/library	MIT License	Credited to the original repository and paper; used for source-scope and bridge/native decision-quality comparisons. Upstream notices are preserved when the package is installed or referenced.
PredOpt-family benchmark assets	Code/benchmark repository	MIT License for the public benchmark repository checked in the audit	Used for shortest-path and knapsack source-scope comparisons; upstream notices are preserved and no stronger SCOPE-routing claim is inferred from native PredOpt evidence.
FRONTIERCO dataset card	Dataset/benchmark card	Dataset-card license marker recorded in the manifest	Used as source-scoped external benchmark evidence; no raw dataset is relicensed by this submission and no deployment-style or same-contract SCOPE routing claim is inferred.
FRONTIERCO code repository	Code repository	No clear reusable software license identified in the repository front matter during the audit	Not vendored, mirrored, or required by the default reviewer path; only paper/repository citation and independently authored diagnostic wrappers or aggregate audit summaries are included.
CO-Bench dataset card	Dataset card	License marked unknown on the dataset card during the audit	Treated conservatively as source-bound; raw upstream data are not redistributed, automatically downloaded, or relicensed. The artifact includes only source citations, Claim Cards, and derived aggregate source-scope records.
CO-Bench code repository	Code repository	No clear reusable software license identified in repository front matter during the audit	Not vendored, mirrored, or required by the default reviewer path; solver-interface checks use independently authored review code and derived summaries rather than copied upstream code.
DATASP code repository	Code repository	No clear reusable software license identified in repository front matter during the audit	Not vendored, mirrored, or required by the default reviewer path; used only for paper/repository citation, source-scope interpretation, and independently generated bridge summaries.
DATASP-associated Warcraft/Cabspotting files	Upstream datasets	Governed by original upstream terms	Not packaged, mirrored, or automatically downloaded; the artifact reports only source-scope Claim Cards and aggregate bridge evidence that do not redistribute the upstream files.

I.5 COMPUTE RESOURCES

The submitted artifact is designed for a CPU-only, offline reviewer path. The default `./run_all.sh` analysis runs from local evidence packs without GPU training, full CAP-DR retraining, network access, external solver-heavy sweeps, or third-party code/data with unknown redistribution terms. It validates locally executable evidence packs, reanalyzes annotation and held-out-study records, rebuilds public source-scope summaries and derived CO-Bench RCSP solver-interface records, and regenerates reports, manifests, and verifier outputs on a standard CPU worker. In our artifact quickstart, the default path normally completes in about one to two minutes. The largest reanalyzed locked object is the RouteScope-Lite request table with 22,500 policy-row records, so the default path has modest CPU-memory requirements. Heavier historical experiments that require GPU training, heavyweight solvers, long wall-clock budgets, external repositories, or upstream assets

with unclear reusable licenses are included only as cached or derived summaries for auditability and are explicitly not downloaded, redistributed, or rerun silently by the default reviewer script.

I.6 FAILURE SEMANTICS

A failed manifest or snapshot check invalidates adjudication-record closure and the claims that depend on it. A failed telemetry-schema check blocks every claim requiring the missing field. A failed registry rebuild blocks the closed-registry count. A failed CLAIMSCOPE-120 analysis blocks fixed-packet agreement claims; a failed reviewer-comparison or construction-audit analysis blocks only the corresponding extension-stage conclusions. A failed ROUTESCOPE-LITE cached reanalysis blocks same-contract conclusion-change claims. Derived-only historical evidence is reported with its historical provenance.

I.7 REVIEWER CHECKS

Table 32 summarizes the versioned reviewer checks. These checks regenerate analyses from their declared raw or cached inputs; they do not imply model retraining or community adoption.

Table 32: Versioned artifact checks and execution modes.

Check	Snapshot / mode	Output artifact
Manifest and snapshot validation	Both / fresh	Hash and chronology report.
Registry and main-table generation	Reviewed / fresh	Registry, exclusions, tables, and CSV mirrors.
Claim Cards, schema, adapters	Reviewed / fresh	Cards plus schema and 15-adapter compliance reports.
ROUTESCOPE-LITE analysis	Reviewed / cached reanalysis	22,500 locked policy-row records, full 15-row table, uncertainty and selection reports.
CLAIMSCOPE-120 analysis	Reviewed / fresh	120 cards, raw-label matrix, tie policies, clustered intervals, and disagreement taxonomy.
Reviewer comparison	Extension / fresh	240 judgments, 36 resolved claim summaries, timing and clustered analyses.
Object construction audit	Extension / fresh	24 paired encodings, field agreements, gate Jaccard, and resolution log.
Public source-scope audit	Reviewed / mixed	Claim Cards, shared gate trace, six source-scope rows, and RCSP solver-interface summaries.