ROUTE-B2B: EVIDENCE ROUTING FOR BUYER QUALIFICATION UNDER INCOMPLETE INFORMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Global business-to-business (B2B) buyer discovery must establish qualification from unevenly observable public information while preserving missing facts as unresolved. ROUTE-B2B connects evidence acquisition, document interpretation, and buyer qualification. Partially Identified Evidence Routing (PIER) maintains qualification bounds and acquires evidence under a fixed budget, returning contact, exclude, or verify according to explicit proof obligations. The Constraint-Compatibility Interface Router (CCIR) uses programmatic evidence contracts to select free-span extraction, source-anchor selection, or agreement-based abstention. Controlled evaluations span four domains and five languages, with paired rich/sparse views preserving candidate facts. Across three frozen open-weight backbones, an exploratory routing replay finds 9.84–20.86% lower source cost for PIER than coverage acquisition, with equal or higher observed correct-resolution rates. CCIR raises exact-card F1 by 30.89 percentage points over free extraction and reduces contract-inexact admission from 21.40% to 2.53%; compatibility relative to fixed anchors varies by backbone. In a compact hosted-system comparison, PIER matches the highest common-readout resolution rate among compared methods at lower source cost than its tied comparators, using a deterministic controller beyond shared extraction. Together, these controlled results support decision-directed evidence economy and expose the distinct roles of contract fidelity, decision coverage, and acquisition cost. Commercial intent remains a separate target for prospective validation.

Code & data | **Project page**

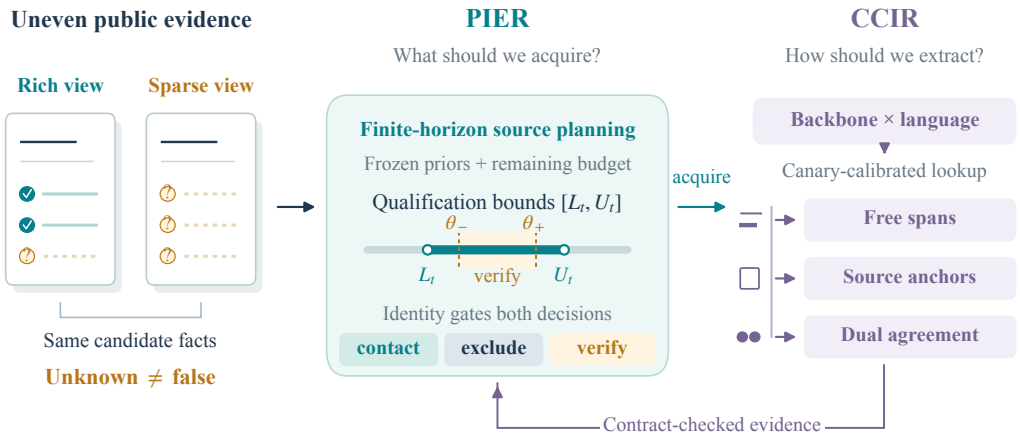


Figure 1: **From incomplete public evidence to an explicit qualification decision.** PIER routes acquisition; CCIR routes interpretation. Admitted cards update qualification bounds, while missing facts remain unresolved. Rich/sparse views preserve latent facts.

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.

1 INTRODUCTION

Business-to-business (B2B) methods discover customers through transaction networks (Jang et al., 2025) and prioritize leads through profiling (Wu et al., 2026a). Market entry also requires supporting evidence: a promising firm may have unverified authorization or service capability amid incomplete market visibility (Vähämäki et al., 2026). Qualification thus asks *which requirements are established, and which missing facts could change the decision*.

This creates two coupled bottlenecks: acquiring information that can change qualification, and interpreting it faithfully enough to justify that change. Missing records remain unresolved, following the positive–unlabeled distinction (Elkan & Noto, 2008) and partial identification (Manski, 2003). ROUTE-B2B operationalizes this curation agenda (Nam et al., 2026): PIER plans acquisitions and stopping from qualification bounds; CCIR calibrates extraction interfaces for exact source witnesses. Together they connect evidence cost, interpretation, and decision sufficiency (Figure 1).

To test routing across qualification blockers, we vary licensing for cultural intellectual-property (IP) goods, intended-use claims for beauty products/devices, import authorization for alcoholic beverages, and service obligations for programmable power systems. All four domains share role, market, and channel requirements (Table 1), separating common reasoning from sector-specific evidence.

Contributions. We contribute (i) provenance-gated bounds with budgeted acquisition and stopping; (ii) contract-calibrated interface selection, tested on 6,720 outputs across three backbones; and (iii) paired visibility controls and common/native readouts against four agentic methods. These controls separate acquisition, extraction, and decision-level effects under shared budgets.

2 RELATED WORK

B2B matching and decision support. Network similarity supports new-market recommendation (Jang et al., 2025); profiling and segmentation support lead ranking and cost-sensitive opportunity forecasting (Wu et al., 2026a; Vähämäki et al., 2026). Co-branding recommendation also integrates partner exploration and budget allocation (Dai et al., 2025). These advances establish predictive value from observed relationships and profiles. For document-based qualification, however, a conversion score leaves open whether a particular requirement has supporting evidence or simply remains unobserved. Conversational rule reasoning addresses a related gap by seeking clarification (Saeidi et al., 2018). We develop this requirement-level perspective into budgeted source acquisition, preserving missing facts as explicit verification needs rather than treating visibility as qualification.

Sequential evidence acquisition. Information value and adaptive optimization connect observations to decision consequences (Howard, 1966; Golovin et al., 2010; Golovin & Krause, 2011). Retrieval-augmented generation (RAG) improves grounded answers through adaptive retrieval (Jiang et al., 2023; Asai et al., 2024); Search-R1 and ReSeek learn search behavior (Jin et al., 2025; Li et al., 2026). Cost-aware RAG further shows that larger evidence budgets can fail to improve answer quality (Wu et al., 2026b). Qualification sharpens this allocation problem: relevant facts have value through their effect on unresolved requirements, and complementary sources may matter only jointly. PIER therefore plans threshold resolution and stopping, rather than fact coverage alone. Table 2 compares planning and verification strategies, including Fleet of Agents (FoA).

Routing, abstention, and evaluation units. RouteLLM uses response preferences to allocate requests across models (Ong et al., 2025). Evidence admission adds an exact claim–witness requirement that preference targets leave unspecified. Even within one backbone, span generation and source-anchor selection impose different output constraints; CCIR calibrates this choice against evidence contracts. Selective prediction trades coverage for risk (Geifman & El-Yaniv, 2017); CCIR applies abstention to evidence admission. Qualification bounds express logical feasibility, distinct from conformal coverage (Angelopoulos & Bates, 2023). Web benchmarks evaluate traversal and synthesis (Wu et al., 2025; Du et al., 2025); attributing qualification gains additionally requires separating acquired facts, extraction fidelity, and final readout. Paired views, exact contracts, and common/native readouts isolate these stages under matched evidence budgets.

3 QUALIFICATION FROM INCOMPLETE EVIDENCE

Rule-based clarification (Saeidi et al., 2018) and context-sufficiency analysis (Joren et al., 2025) motivate separating missing evidence from interpretation errors. For buyer curation (Nam et al., 2026), define $q = (o, m, \tau, \mathcal{P}, B)$: offering, market, cutoff, proof obligations $\mathcal{P} = \{p_k\}_{k=1}^{K}$, and candidate budget. Provenance-linked cards (Moreau & Missier, 2013) record entity, predicate, support/contradiction, source, authority, language, publication/retrieval times, document hash, and immutable span. Admission checks predicate, span, time, and authority; temporal reasoning failures (Zhou et al., 2026) reinforce cutoff checks. Semantic errors can still affect qualification.

Motivated by set-valued partial identification (Manski, 2003), we retain every compatible assignment of latent satisfaction $z_k \in \{0, 1\}$, with declared weights $w_k \geq 0$, $\sum_k w_k = 1$. Support fixes $[\ell_k, u_k] = [1, 1]$; authoritative contradiction fixes $[0, 0]$; absence or unresolved conflict leaves $[0, 1]$. Our logical bounds are

$$\mathcal{Z}(E) = \{z \in \{0,1\}^K : \ell_k(E) \leq z_k \leq u_k(E)\ \forall k\},$$
$$[L(E), U(E)] = \left[\sum_k w_k \ell_k(E),\ \sum_k w_k u_k(E)\right]. \tag{1}$$

These bounds are sharp over the rectangular feasible set; an unknown license retains both possibilities. Selective prediction (Geifman & El-Yaniv, 2017) and recent LLM abstention evaluation (Kirichenko et al., 2025) motivate abstention; our rule uses identity gate $I(E)$ and thresholds $\theta_- < \theta_+$:

$$d(E) = \begin{cases} \mathsf{contact}, & I(E) = 1 \text{ and } L(E) \geq \theta_+, \\ \mathsf{exclude}, & I(E) = 1 \text{ and } U(E) < \theta_-, \\ \mathsf{verify}, & \text{otherwise.} \end{cases} \tag{2}$$

Both resolved actions require identity; scores in the threshold gap map to verify. Weights and thresholds ($\theta_- = 0.35$, $\theta_+ = 0.72$) are declared before evaluation. Compliance certification additionally requires jurisdiction-specific hard constraints and validation against the intended market's applicable qualification requirements.

Proposition 1 (Conditional soundness)**.** *If each resolved predicate bound contains its true value and the identity gate is correct, then $L(E) \leq \sum_k w_k z_k \leq U(E)$, and every resolved decision in Equation 2 agrees with the same rule applied to complete information. Adding correct, consistent evidence monotonically refines the feasible interval of qualification scores.*

The coordinatewise proof (Appendix B) relies on valid evidence, independently of whether missingness is ignorable (Rubin, 1976). Incorrect admission can violate this premise; conflicting evidence reopens affected bounds for further source-level verification.

4 TWO-LAYER EVIDENCE ROUTING

4.1 PIER: PLAN FOR RESOLUTION, THEN AMBIGUITY, THEN COST

Following costly information acquisition (Howard, 1966; Golovin et al., 2010), action a reveals bundle outcome X_a at cost $c_a > 0$. We define decision ambiguity by

$$A(E) = \mathbf{1}\{d(E) = \mathsf{verify}\}\left([\min\{U(E), \theta_+\} - \max\{L(E), \theta_-\}]_+ + \lambda_v\right), \qquad \lambda_v = 0.25. \tag{3}$$

The overlap measures width inside the decision band; λ_v penalizes verify at zero overlap. Greedy-Cov (greedy evidence coverage) maximizes expected weighted predicate coverage per cost; one-step ambiguity maximizes $(A(E) - \mathbb{E}[A(E \cup X_a)])/c_a$. Greedy guarantees require suitable structure (Golovin & Krause, 2011); here, complementary sources can jointly cross a qualification threshold, motivating planning beyond immediate evidence gain.

For remaining actions $\mathcal{R}$ and budget b, PIER uses dynamic programming (Bellman, 1957) with lexicographic terminal value $\mathbf{g}(E) = (\mathbf{1}\{d(E) \neq \mathsf{verify}\}, -A(E), 0)$:

$$\mathbf{J}(E, \mathcal{R}, b) = \operatorname{lexmax}\left\{\mathbf{g}(E),\ \operatorname*{lex\,max}_{\substack{a \in \mathcal{R} \\ c_a \leq b}} \mathbf{Q}(E, \mathcal{R}, b, a)\right\},$$
$$\mathbf{Q}(E, \mathcal{R}, b, a) = \mathbb{E}_{X_a}[\mathbf{J}(E \cup X_a, \mathcal{R} \setminus \{a\}, b - c_a)] - (0, 0, c_a). \tag{4}$$

The value vector prioritizes expected resolution, ambiguity, then cost. *Horizon*: Full evaluates all budget-feasible continuations; One replans one step after each observation. STOP On lets $\mathbf{g}(E)$ compete with acquisition; Off continues until resolution or no affordable action remains. Factorized outcome priors and memoization enable exact five-bundle planning; complete-information labels assess realized correctness under the same qualification rule.

The open-weight replay uses simulator-supplied truth-frequency priors and distractor roles. The hosted study uses opaque source IDs and a common $(0.4, 0.4, 0.2)$ support/contradiction/unresolved prior. GreedyCov acquires while unresolved with affordable sources; PIER also evaluates stopping. Savings compare both acquisition and stopping under the shared evidence budget.

4.2 CCIR: CONTRACT-CALIBRATED EXTRACTION INTERFACES

To address the interpretation bottleneck, an *interface* defines the output: Free generates a state and span; Anchor selects a state and opaque identifier mapped to a span. CCIR calibrates the choice, with dual agreement as fallback, holding backbone and admission checks fixed. In tables, *Policy* denotes the extraction policy; *Router* denotes the source-acquisition policy.

For calibration (canary) example i and interface $h \in \{F, A\}$, let C_{ih} indicate exact contract correctness (including a valid unresolved response when no card is expected), and W_{ih} indicate an incorrect admitted card. CCIR compares paired utilities:

$$u_{ih} = C_{ih} - W_{ih}, \qquad \widehat{\Delta}_{m,\ell} = \frac{1}{n_\ell} \sum_{i:\mathrm{lang}(i)=\ell} (u_{iA} - u_{iF}), \qquad n_\ell = |\mathcal{C}_\ell|. \tag{5}$$

Each language's calibration set $\mathcal{C}_\ell$ has 3 claim states $\times$ 2 surface families $\times$ 4 variants: 24 examples, 120 across five languages. Within backbone m, exact paired sign-randomization (Dror et al., 2018) requires Holm-adjusted $p \leq 0.10$ (Holm, 1979) across languages and $|\widehat{\Delta}_{m,\ell}| \geq 0.05$. Pooled-backbone backoff requires $p \leq 0.05$ with the same effect threshold. Otherwise, *dual agreement* admits matching states and mapped spans, using two calls per request to corroborate admitted evidence before it reaches the downstream qualification controller.

Behavior-targeted testing (Ribeiro et al., 2020) motivates our canary construction. Canaries are disjoint from evaluation in wording, entities, layouts, seed, and hashes. Selection is frozen, with backbone parameters unchanged. CCIR combines paired contract calibration, hierarchical backoff, and agreement-gated admission using programmatic labels.

5 CONTROLLED BENCHMARK AND COMPARISON PROTOCOL

5.1 CONSTRUCTION AND INDEPENDENT UNITS

Practitioners from each domain reviewed the earlier candidate rubric for evidence sufficiency, actions, and verification responsibility. The benchmark operationalizes this practitioner-informed curation agenda (Nam et al., 2026) through programmatic instance labels (Table 1). Each of 80 latent profiles (20/domain) has six weighted Boolean obligations and paired rich/sparse views exposing two/zero predicates. Behavior-targeted pairs (Ribeiro et al., 2020) fix facts, language, and surface, varying visibility and distractors. Six fact-bearing documents and one neutral or unauthorized-negative distractor yield seven extraction requests per episode. English, Korean, French, German, and Japanese are balanced. *Document surface* denotes rendering format: narrative prose, compact field registries, announcement-style bulletins, or table-like rows; all are balanced. Organizations are fictional; addresses use `.invalid`. The evidence cutoff is 31 December 2025, shared by all candidate profiles and paired views for checking publication-time eligibility of source documents.

Programmatic gold labels encode qualification under explicit evidence requirements; a shared acquisition graph supports matched domain comparisons. Paired inference retains both views in each *profile cluster*, preserving within-unit dependence (Cameron & Miller, 2015).

5.2 OPEN-WEIGHT AND HOSTED-SYSTEM CONDITIONS

Sealed open-weight study (S2). Qwen3.5-4B, Gemma 4 E4B-IT, and Phi-4-mini-instruct (Qwen Team, 2026; Gemma Team, 2026; Microsoft et al., 2025) each run 1,120 requests per interface:

Table 1: **Shared decision structure, professionally grounded evidence roles.** Cited sources motivate illustrative evidence roles; all domains also specify role, category, market, and channel.

Domain	Capability evidence	Fulfillment evidence
Cultural-IP goods	Licensing and quality control (McKay, 2015)	Territorial distribution (McKay, 2015)
Beauty products/devices	Intended use and claims (FDA, n.d.b; FDA, n.d.a)	Returns and replacement (Medicube, n.d.)
Alcoholic beverages	Import/label authorization (TTB, 2025)	Excise-movement records (European Commission, 2023)
Programmable power	Calibration and service (AMETEK, 2024)	Commissioning and support (AMETEK, 2024)

Table 2: **Recent comparators on one paid catalog.** Asterisks denote task/budget adaptations evaluated with the same hosted system, detailed in Appendix D.

Comparator (reference)	Acquisition mechanism
ToolTree* (Yang et al., 2026)	Tree search, pre/post evaluation, bidirectional pruning
ATLAS* (Choi et al., 2026)	Constraint manager, planner, verifier, gap search
Fleet of Agents* (Klein et al., 2025)	Multiple trajectories, heuristic values, resampling
MindSearch* (Chen et al., 2025)	Graph planner, search worker, result integration
Direct agent	Next-source or stop decision from current state
GreedyCov	Expected unresolved-predicate coverage per cost

6,720 evaluation and 720 separate canary outputs. Inference uses NVIDIA A100 graphics processing units (GPUs), 16-bit brain floating point (BF16), greedy decoding, batch size one, and 256 maximum new tokens, with software and canary selection frozen (Appendix C).

Eight policies reuse outputs (Table 4): Free, Anchor, CCIR, four selector ablations, and a nondeployable test-label oracle. Fixed-policy post-hoc replay with GreedyCov and PIER retains all 48 configurations (42 deployable; Appendix E).

Compact hosted comparison. A metadata-only selection of two profiles per domain–language cell yields 80 episodes / 40 profiles. Seven controllers share a 560-request Anchor cache, requesting `gpt-6-astra` with effort `ultra` through isolated Codex command-line interface (CLI) sessions. Server-effective effort and hidden computation remain provider-controlled. Inputs exclude gold, other trajectories, and unacquired sources. Every method has budget $B = 2.45$; Direct is capped at six logical requests per episode and the four adapted agents at 16. Global charging includes acquisitions on discarded branches as well as the retained final trajectory.

Table 2 summarizes mechanisms; Appendix D details task adaptations.

The *common readout* applies Equation 2 to initial and globally acquired evidence, including discarded branches. The *native readout* scores the retained trajectory's terminal answer. With extraction and prices fixed, contrasts reflect both terminal readout and branch-level evidence aggregation.

5.3 Metrics and statistical separation

For N episodes, let y_i be the complete-information label, $\hat{y}_i$ the prediction, and $r_i = \mathbf{1}\{\hat{y}_i \neq \mathsf{verify}\}$. Coverage and selective accuracy (SA) (Geifman & El-Yaniv, 2017) give the joint yield, correctly resolved decision rate (CRDR):

$$\mathrm{CRDR} = \frac{\sum_i r_i \mathbf{1}\{\hat{y}_i = y_i\}}{N}, \qquad \mathrm{Cov} = \frac{\sum_i r_i}{N}, \qquad \mathrm{SA} = \frac{\sum_i r_i \mathbf{1}\{\hat{y}_i = y_i\}}{\sum_i r_i}. \tag{6}$$

Err. is incorrect resolution / N; qualified/unsafe exclusion condition on $y_i = \mathsf{contact}$ / $y_i \neq \mathsf{exclude}$. Undefined ratios are not applicable (NA). Cost is mean bundle cost/episode in our designed units (Wu et al., 2026b); initial bundles cost zero. Calls count S2 initial/acquired-document extraction (two for dual), or hosted controller requests excluding shared extraction.

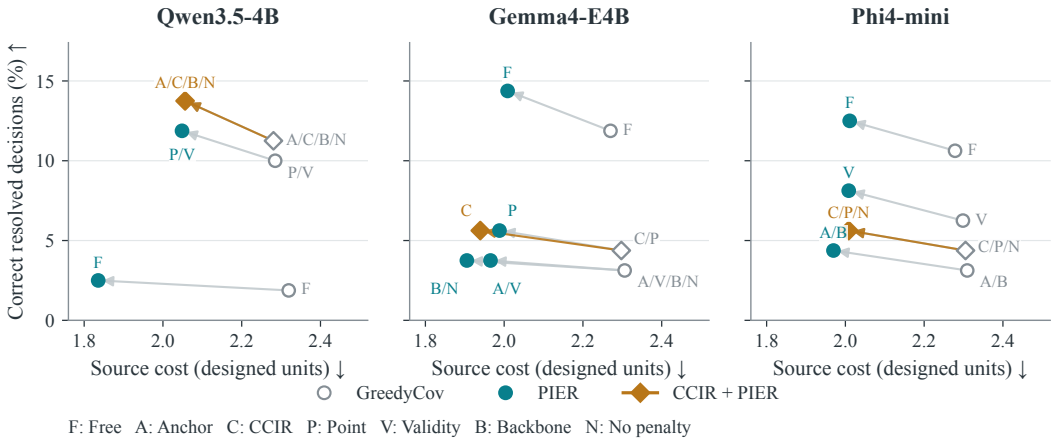


Figure 2: **All 42 deployable S2 configurations at one budget.** Each backbone pairs seven extraction policies with two routers; coincident points share labels. Four-axis Pareto comparisons additionally include errors and extraction calls, evaluated for each deployable configuration.

Table 3: **System composition and matched component ablation.** Three-backbone means; CRDR ↑/Err. ↓ (% of episodes), Cost/Calls ↓ per episode. GreedyCov is a greedy acquisition baseline; CCIR is fixed in (b). Tint/bold mark proposed rows; †: lower resources than matched STOP-off in each backbone (45-test Holm $p = .00225$). Panel (a) is descriptive; full results: Appendix F.

(a) Extraction × acquisition

Policy	Router	CRDR ↑	Cost ↓	Calls ↓	Err. ↓
Free	GreedyCov	8.12	2.289	5.223	0.42
Free	PIER	9.79	1.952	4.365	0.21
Anchor	GreedyCov	5.83	2.299	5.248	0.00
Anchor	PIER	7.29	1.997	4.419	0.00
CCIR	GreedyCov	6.67	2.294	6.637	0.00
CCIR	**PIER**	**8.33**	**2.001**	**5.567**	**0.00**

(b) Horizon × STOP

Router	Horizon	STOP	CRDR ↑	Cost ↓	Calls ↓
GreedyCov	–	–	6.67	2.294	6.637
PIER	Full	Off	8.33	2.257	6.517
PIER	**Full**	**On**	**8.33**	**2.001**†	**5.567**†
PIER	One	Off	8.12	2.248	6.617
PIER	One	On	8.12	**1.744**†	**5.096**†

Following evidence-linked evaluation (Thorne et al., 2018; Gao et al., 2023), our exact-card precision, recall, and harmonic-mean F1 require correct state and full span. Inexact/State acc. are incorrect admissions/correct states per request; S2 weights backbones equally. Our sealed joint criterion requires at least +3 percentage points (pp) F1 over *both* fixed interfaces, positive 95% lower bounds, Holm-adjusted correctness $p \leq .05$, and an incorrect-admission upper bound of at most +1 pp. F1 intervals and correctness tests retain distinct inferential targets (Dror et al., 2018).

Confidence intervals (CIs) use 10,000 domain-stratified paired profile-cluster bootstrap draws (Cameron & Miller, 2015), keeping views and methods together. Separate Holm families cover two S2 correctness, 171 exploratory replay, and six hosted CRDR tests (Holm, 1979). Cost/risk intervals are nominal; Appendix E details paired tests and their assumptions.

6 Results: Acquisition Economy and Interface Compatibility

6.1 PIER lowers acquisition cost under matched extraction policies

With extraction fixed, PIER uses 9.84–20.86% less source cost than GreedyCov across 21 deployable backbone–policy pairs, with nonnegative observed CRDR differences. All cost contrasts have Holm-adjusted (Holm, 1979) $p = 0.00855$ in the exploratory 171-test family. Figure 2 shows all operating points; Table 3 separates system composition from CCIR-fixed planning/stopping. Every matched pair empirically dominates GreedyCov in CRDR, errors, source cost, and extraction calls; adjusted tests support lower cost within this declared family of exploratory comparisons.

Table 4: **Fixed interfaces and CCIR component controls.** F1/Inexact $\times 100$; subscripts are profile-bootstrap standard errors (pp). Macro equally weights backbones. Calls/request excludes calibration; Oracle* uses test labels and is nondeployable.

Policy	Exact-card F1 ↑				Inexact ↓	Calls ↓
	Qwen (Qwen Team, 2026)	Gemma (Gemma Team, 2026)	Phi (Microsoft et al., 2025)	Macro Equal weight		
Free	$27.20_{\pm 4.88}$	$54.10_{\pm 4.71}$	$14.93_{\pm 3.14}$	$32.08_{\pm 3.29}$	$21.40_{\pm 2.31}$	1.000
Anchor	$88.00_{\pm 1.97}$	$64.37_{\pm 3.49}$	$40.16_{\pm 2.91}$	$64.17_{\pm 2.39}$	$1.79_{\pm 0.28}$	1.000
CCIR	$\mathbf{88.00}_{\pm 1.97}$	$\mathbf{55.34}_{\pm 4.50}$	$\mathbf{45.55}_{\pm 2.81}$	$\mathbf{62.96}_{\pm 2.67}$	$\mathbf{2.53}_{\pm 0.40}$	**1.267**
Point	$80.86_{\pm 2.95}$	$67.21_{\pm 3.63}$	$45.55_{\pm 2.81}$	$64.54_{\pm 2.69}$	$3.42_{\pm 0.74}$	1.000
Validity	$80.86_{\pm 2.95}$	$64.37_{\pm 3.49}$	$29.96_{\pm 3.47}$	$58.40_{\pm 2.37}$	$7.98_{\pm 1.25}$	1.000
Backbone	$88.00_{\pm 1.97}$	$51.05_{\pm 4.40}$	$40.16_{\pm 2.91}$	$59.74_{\pm 2.59}$	$1.70_{\pm 0.28}$	1.333
No penalty	$88.00_{\pm 1.97}$	$51.05_{\pm 4.40}$	$45.55_{\pm 2.81}$	$61.53_{\pm 2.61}$	$2.08_{\pm 0.29}$	1.333
Oracle*	$88.00_{\pm 1.97}$	$67.21_{\pm 3.63}$	$45.55_{\pm 2.81}$	$66.92_{\pm 2.44}$	$2.62_{\pm 0.40}$	1.000

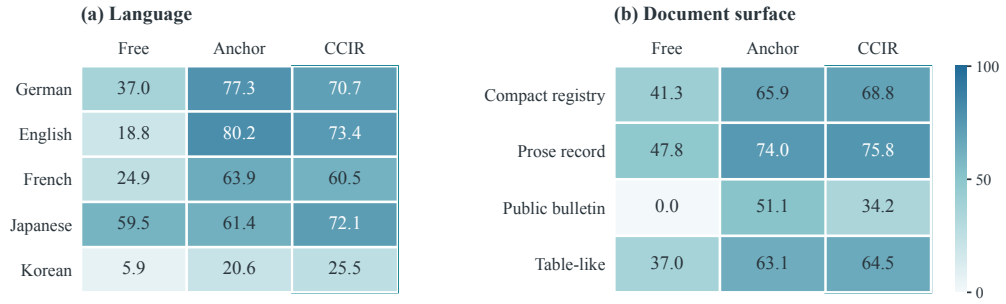


Figure 3: **Language and surface shape compatibility.** Exact-card F1, equally averaged over three backbones; all levels shown. Shared-profile diagnostics are descriptive.

For CCIR, CRDR rises from 6.67 to 8.33% alongside lower cost and calls (Table 3(a)); CRDR-difference intervals cross zero (Table 10). CCIR+PIER lies on every backbone's empirical four-axis Pareto set. Cost–call tradeoffs (Appendix E) preserve application-dependent choices, consistent with context-dependent selector rankings in cost-aware RAG (Wu et al., 2026b).

6.2 CCIR IMPROVES ON FREE EXTRACTION, WITH BACKBONE-DEPENDENT COMPATIBILITY

Table 4 isolates interpretation across eight policies. CCIR increases mean exact-card F1 from 32.08 to 62.96 and lowers incorrect admission from 21.40% to 2.53% relative to Free. The paired F1 gain is $+30.89$ pp [26.64,35.24], with correctness-test Holm $p = 2.0 \times 10^{-5}$. Against Anchor, the F1 difference is -1.21 pp [-3.97,1.53], while the incorrect-admission upper bound is $+1.49$ pp, exceeding the prespecified $+1$ pp margin. Confirmatory support therefore applies to the Free contrast; the original joint two-interface criterion remains unmet (Figure 6).

Qwen selects Anchor in every language (88.00 F1). Phi selects Free for Japanese/Korean and Anchor elsewhere (45.55 versus Anchor's 40.16 F1). Gemma uses dual agreement in four languages (1.80 calls/request): 55.34 F1 versus Anchor's 64.37, 96.13% admitted-card precision, and 38.85% recall. This agreement tradeoff parallels selective prediction (Geifman & El-Yaniv, 2017); coincident policies reuse the same outputs and therefore contribute identical evidence cards to qualification.

Figure 3 exposes heterogeneity hidden by pooled F1: CCIR exceeds Anchor in Japanese and Korean, whereas Anchor leads in the other languages; document formatting also changes the compatibility ordering among interfaces on the same evaluation requests.

Of 308 admitted Free bulletin cards, 304 quote gold substrings and 297 have correct states; all miss the full witness. F1 scores exact recovery; CRDR scores qualification. Evidence and entailment evaluation (Thorne et al., 2018; Gao et al., 2023) motivate assessing task-sufficient partial spans.

Table 5: **Hosted comparison: two readouts from the same runs.** 80 episodes / 40 profiles; rates in %, Cost and controller Calls per episode. Asterisks mark task/budget adaptations. All methods use the same separately charged 560-request extraction cache.

Method	Common readout			Native readout			Resources	
	CRDR	Cov.	Sel. acc.	CRDR	Cov.	Sel. acc.	Cost ↓	Calls ↓
GreedyCov	17.50	17.50	100.00	17.50	17.50	100.00	2.276	0.00
PIER	**17.50**	**17.50**	**100.00**	**17.50**	**17.50**	**100.00**	**2.153**	**0.00**
Direct Astra	12.50	12.50	100.00	32.50	33.75	96.30	1.916	3.33
ATLAS* (Choi et al., 2026)	11.25	11.25	100.00	5.00	5.00	100.00	1.636	13.03
MindSearch* (Chen et al., 2025)	17.50	17.50	100.00	6.25	6.25	100.00	2.263	14.99
ToolTree* (Yang et al., 2026)	8.75	8.75	100.00	17.50	18.75	93.33	1.788	13.59
FoA* (Klein et al., 2025)	12.50	12.50	100.00	18.75	20.00	93.75	1.674	9.31

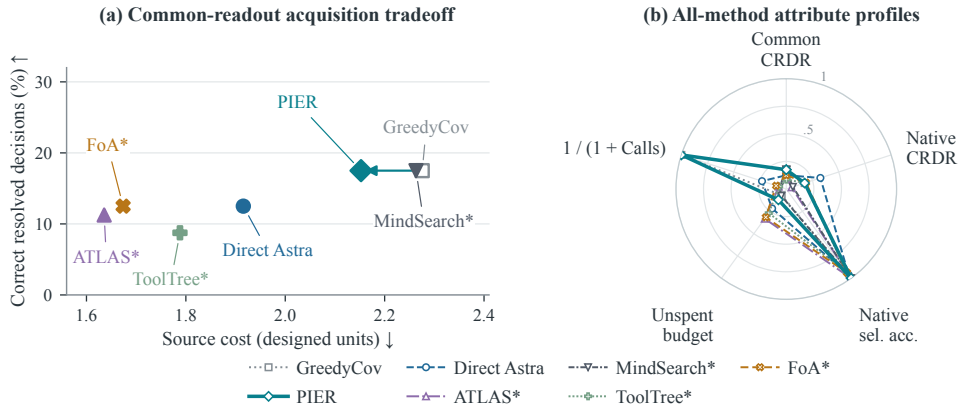


Figure 4: **Hosted baseline tradeoffs with Anchor extraction fixed.** All seven methods from Table 5: common-readout cost–resolution (left) and fixed-scale attribute profiles (right). Source cost measures designed acquisitions. The radar retains native resolution and selective accuracy alongside resource consumption from the same completed hosted runs.

6.3 A STRONG HOSTED SYSTEM RETAINS THE ACQUISITION TRADEOFF

With Anchor fixed, PIER, GreedyCov, and MindSearch* (Chen et al., 2025) tie at 17.5% common CRDR (Table 5). PIER saves 5.41% source cost versus GreedyCov and 4.89% versus MindSearch*, with paired differences -0.123 $[-0.151, -0.097]$ and -0.111 $[-0.141, -0.078]$. Its deterministic controller adds zero requests beyond shared extraction, versus MindSearch*'s 15.0 per episode.

All six common-CRDR tests have Holm $p \geq 0.20069$: quality differences are descriptive. Direct, ATLAS*, ToolTree*, and FoA* cost less but resolve fewer common-readout episodes. Native Direct instead reaches 32.5% CRDR and 96.30% selective accuracy; FoA* reaches 18.75% CRDR with one incorrect resolution. Figure 4 retains all methods and both readouts, making their decision–resource tradeoffs explicit under shared acquisition and extraction conditions.

7 WHAT THE TRADEOFFS IMPLY

Connecting contract fidelity to decision value. Figure 5 links witness recovery to the six systems in Table 3(a). In information-value terms (Howard, 1966), a predicate aids resolution when its evidence can move a bound across a threshold. Exact-card F1 scores witness recovery, whereas decision yield depends on acquisition paths and predicate weights. These targets can rank systems differently: Free+PIER achieves higher observed resolution, while CCIR recovers more exact witnesses than Free. CCIR+PIER couples witness fidelity with budgeted acquisition. Error counts (one versus zero) motivate larger risk-focused cohorts for precise safety comparisons.

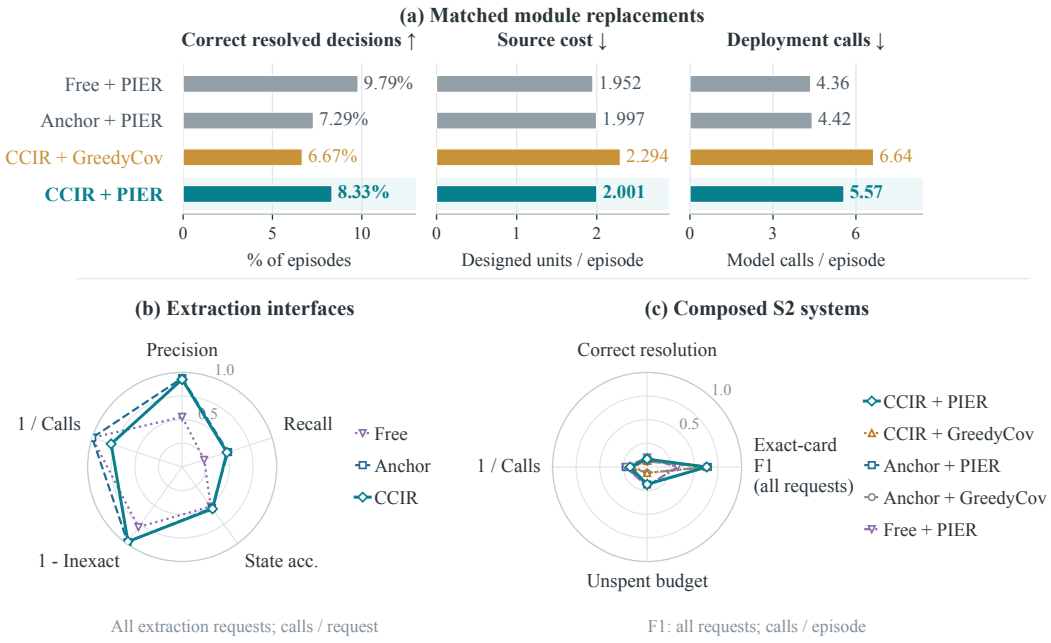


Figure 5: **Extraction attributes, composition, and actual resource use.** S2 three-backbone means: (a) module replacements in original units; (b) extraction and (c) joint-system radars on fixed $[0, 1]$ axes. Resource axes use $1 - \text{Cost}/2.45$ and $1/\text{Calls}$ (per request in b, per episode in c). F1 covers all requests; all six systems appear in Table 3(a).

Stopping identifies the clearest acquisition mechanism. With CCIR fixed, PIER saves 12.77% source cost and 16.13% calls relative to GreedyCov. The matched horizon×STOP replay (Table 3(b)) isolates value-based stopping: at full horizon, cost falls **9.70–13.17%** and calls **12.32–16.15%**, with all 480 paired decisions unchanged and every resource contrast at Holm $p = .00225$. One-step with STOP costs less still; full planning adds 0.625 pp CRDR only for Qwen ($p_H = 1$). Decision-directed stopping thus provides the clearest component gain, linking information value (Howard, 1966) to cost-aware evidence stopping (Wu et al., 2026b); Appendix F retains all 45 exploratory tests with their paired contrasts and adjusted significance levels.

From uncertainty preservation to broader access. Paired sparse views pinpoint evidence-access bottlenecks (Figure 8). Verify motivates clarification when predicates remain unknown (Saeidi et al., 2018); fully known scores in the threshold gap instead remain unresolved by design. Transfer tests should vary source availability, witness boundaries, and prior calibration under fixed obligations, then compare coverage and errors across languages at matched budgets using risk–coverage analysis (Geifman & El-Yaniv, 2017). Original-language witnesses and provenance support such audits. Prospective field studies can then assess the relationship between evidence-supported qualification, buyer intent, and realized transactions in each target market.

8 Conclusion

ROUTE-B2B connects candidate investigation to evidence-supported qualification. The studies support PIER's evidence economy, with optional stopping its clearest measured component. CCIR improves exact-witness recovery over Free, with backbone-dependent tradeoffs against Anchor. These findings motivate a practical evaluation principle: assess evidence fidelity, decision resolution, and resources as distinct outcomes. Paired views, fixed contracts, and common/native readouts make that separation reproducible, providing an executable foundation for public-document and commercial validation. The modular contracts separate acquisition priorities from witness requirements, supporting domain-specific adaptation under an unchanged, explicit qualification readout.

AI USE STATEMENT

Generative AI assistance was used to refine research questions and protocols, develop the formalization and mathematical arguments, implement methods and programmatic data generators, draft multilingual templates, inspect and interpret results, verify references, and prepare the manuscript, tables, and figures. Evaluated language-model systems are separately specified in Sections 5–6. Qualification gold facts and exact witness contracts are defined by deterministic generation and scoring rules. Practitioner feedback informed task design (Section 5); frozen artifacts, automated checks, and paired analyses support instance-level verification. Deployment validation would extend this design input to representative real firms, public documents, and commercial outcomes. The authors are responsible for the final manuscript and research artifacts.

ETHICS STATEMENT

Generated records and reserved domains protect real firms and private contacts from experimental labeling or outreach. Contact and exclude denote benchmark outcomes; commercial action requires separate authorization and verification. Public-data absence is retained as unresolved evidence. Unequal unresolved rates nevertheless require attention before deployment. A real-source release would need source-specific permission, data minimization, temporal verification, and independently validated qualification constructs; access allowed by a robots file alone is insufficient.

REPRODUCIBILITY STATEMENT

The appendix specifies proof obligations, decision rules, model revisions, source costs, interface calibration, study amendments, statistical units, and complete method comparisons. The research repository retains versioned configurations, frozen manifests, raw-output hashes, request-level records, and scripts that regenerate the reported analyses. Empirical tables and figures are generated directly from the complete machine-readable study summaries. The project page supplies central processing unit (CPU) ablation and table-uncertainty packages alongside the fixed code snapshot. For the hosted system, requested settings, client version, isolation configuration, and operational recovery records provide the reproducibility boundary; provider-controlled weights and hidden execution limit bitwise repetition.

REFERENCES

AMETEK. *Reliability Assurance Program*. AMETEK Programmable Power, 2024. URL https://www.programmablepower.com/service-and-support/reliability-assurance. Service brochure; revision code 20240625; especially pp. 2–3 and 6.

Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023. doi: 10.1561/2200000101. URL https://doi.org/10.1561/2200000101.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. SELF-RAG: Learning to retrieve, generate, and critique through self-reflection. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=hSyW5go0v8.

Richard Bellman. *Dynamic Programming*. Princeton University Press, 1957.

A. Colin Cameron and Douglas L. Miller. A practitioner's guide to cluster-robust inference. *Journal of Human Resources*, 50(2):317–372, 2015. doi: 10.3368/jhr.50.2.317. URL https://jhr.uwpress.org/content/50/2/317.

Zehui Chen, Kuikun Liu, Qiuchen Wang, Jiangning Liu, Wenwei Zhang, Kai Chen, and Feng Zhao. MindSearch: Mimicking human minds elicits deep AI searcher. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/e01c431bbb83153632c0dcfaf8ccda0a-Abstract-Conference.html.

Jihye Choi, Jinsung Yoon, Jiefeng Chen, Somesh Jha, and Tomas Pfister. ATLAS: Constraints-aware multi-agent collaboration for real-world travel planning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=mIYGiBf9Pm.

Xiangxiang Dai, Xiaowei Sun, Jinhang Zuo, Xutong Liu, and John C. S. Lui. A unified online-offline framework for co-branding campaign recommendations. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, volume 2, pp. 427–438, 2025. doi: 10.1145/3711896.3736824.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of ACL*, pp. 1383–1392. Association for Computational Linguistics, 2018. doi: 10.18653/v1/P18-1128. URL https://aclanthology.org/P18-1128/.

Mingxuan Du, Benfeng Xu, Chiwei Zhu, Xiaorui Wang, and Zhendong Mao. Deepresearch bench: A comprehensive benchmark for deep research agents. *arXiv preprint arXiv:2506.11763*, 2025.

Charles Elkan and Keith Noto. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 213–220, 2008. doi: 10.1145/1401890.1401920. URL https://doi.org/10.1145/1401890.1401920.

European Commission. All excise duty transactions in the EU become fully electronic as of 13 february. Directorate-General for Taxation and Customs Union, February 2023. URL https://taxation-customs.ec.europa.eu/news/all-excise-duty-transactions-eu-become-fully-electronic-13-february-2023-02-10_en. News announcement, February 10, 2023.

FDA. Aesthetic (cosmetic) devices. U.S. Food and Drug Administration, n.d.a. URL https://www.fda.gov/medical-devices/products-and-medical-procedures/aesthetic-cosmetic-devices. Accessed September 20, 2026.

FDA. Is it a cosmetic, a drug, or both? (or is it soap?). U.S. Food and Drug Administration, n.d.b. URL https://www.fda.gov/cosmetics/cosmetics-laws-regulations/it-cosmetic-drug-or-both-or-it-soap. Accessed September 20, 2026; intended-use and product-claims classification.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate text with citations. In *Proceedings of EMNLP*, pp. 6465–6488. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.398. URL https://aclanthology.org/2023.emnlp-main.398/.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html.

Gemma Team. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026. URL https://arxiv.org/abs/2607.02770.

Daniel Golovin and Andreas Krause. Adaptive submodularity: Theory and applications in active learning and stochastic optimization. *Journal of Artificial Intelligence Research*, 42:427–486, 2011. doi: 10.1613/jair.3278. URL https://doi.org/10.1613/jair.3278.

Daniel Golovin, Andreas Krause, and Debajyoti Ray. Near-optimal Bayesian active learning with noisy observations. In *Advances in Neural Information Processing Systems*, volume 23, pp. 766–774, 2010. URL https://proceedings.neurips.cc/paper/2010/hash/1e6e0a04d20f50967c64dac2d639a577-Abstract.html.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. URL https://www.jstor.org/stable/4615733.

Ronald A. Howard. Information value theory. *IEEE Transactions on Systems Science and Cybernetics*, 2(1):22–26, 1966. doi: 10.1109/TSSC.1966.300074. URL https://ieeexplore.ieee.org/document/4082064/.

Kabsoo Jang, Jeongsub Choi, Ho-shin Lee, and Byunghoon Kim. A novel similarity-based recommendation for identifying potential customers in new markets using an inter-firm transaction network. *Technological Forecasting and Social Change*, 216:124151, 2025. doi: 10.1016/j.techfore.2025.124151.

Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7969–7992, 2023. doi: 10.18653/v1/2023.emnlp-main.495.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-R1: Training LLMs to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*, 2025.

Hailey Joren, Jianyi Zhang, Chun-Sung Ferng, Da-Cheng Juan, Ankur Taly, and Cyrus Rashtchian. Sufficient context: A new lens on retrieval augmented generation systems. In *International Conference on Learning Representations*, 2025.

Polina Kirichenko, Mark Ibrahim, Kamalika Chaudhuri, and Samuel J. Bell. AbstentionBench: Reasoning LLMs fail on unanswerable questions. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-5729. Datasets and Benchmarks Track.

Lars Henning Klein, Nearchos Potamitis, Roland Aydin, Robert West, Caglar Gulcehre, and Akhil Arora. Fleet of agents: Coordinated problem solving with large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 30986–31019. PMLR, 2025. URL https://proceedings.mlr.press/v267/klein25a.html.

Shiyu Li, Yifan Wang, Peiming Li, Zheng Wei, and Yang Tang. ReSeek: A self-correcting framework for search agents with instructive rewards. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://github.com/TencentBAC/ReSeek.

Charles F. Manski. *Partial Identification of Probability Distributions*. Springer Series in Statistics. Springer, New York, 2003. doi: 10.1007/b97478. URL https://link.springer.com/book/10.1007/b97478.

Kenneth D. McKay. Guide on trademark licensing. Technical Report CDIP/16/INF/2, World Intellectual Property Organization, September 2015. URL https://dacatalogue.wipo.int/projectfiles/DA_10_03/CDIP_16_INF_2/EN/cdip_16_inf_2_guide_tm_licensing.pdf. September 28, 2015; especially Sections 4.3 and 4.4.2.

Medicube. Refund policy. APR Corporation, Medicube U.S. storefront, n.d. URL https://medicube.us/policies/refund-policy. Accessed September 20, 2026; direct-purchase refund and replacement policy.

Microsoft, Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, et al. Phi-4-Mini Technical Report: Compact yet powerful multimodal language models via mixture-of-LoRAs. *arXiv preprint arXiv:2503.01743*, 2025.

Luc Moreau and Paolo Missier. PROV-DM: The PROV data model. W3C recommendation, World Wide Web Consortium, April 2013. URL https://www.w3.org/TR/2013/REC-prov-dm-20130430/.

Jason K. Nam, SangWun Lee, Hyesun Ryu, Jisoo Min, Juheon Yuk, Tae Guen Lee, and Seong Hoon Jeon. Buyer discovery is a data-curation problem: A position on decision-grade evidence for global B2B AI. In *Curated Data for Efficient Learning Workshop at the European Conference on Computer Vision*, 2026. URL https://openreview.net/forum?id=wGmafemQny. Archival workshop paper.

Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E. Gonzalez, M. Waleed Kadous, and Ion Stoica. RouteLLM: Learning to route LLMs from preference data. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/5503a7c69d48a2f86fc00b3dc09de686-Abstract-Conference.html.

Qwen Team. Qwen3.5: Towards native multimodal agents. https://qwen.ai/blog?id=qwen3.5, 2026.

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. doi: 10.18653/v1/2020.acl-main.442.

Donald B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976. doi: 10.1093/biomet/63.3.581. URL https://academic.oup.com/biomet/article-abstract/63/3/581/270932.

Marzieh Saeidi, Max Bartolo, Patrick Lewis, Sameer Singh, Tim Rocktäschel, Mike Sheldon, Guillaume Bouchard, and Sebastian Riedel. Interpretation of natural language rules in conversational machine reading. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2087–2097. Association for Computational Linguistics, 2018. doi: 10.18653/v1/D18-1233. URL https://aclanthology.org/D18-1233/.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of NAACL-HLT*, pp. 809–819. Association for Computational Linguistics, 2018. doi: 10.18653/v1/N18-1074. URL https://aclanthology.org/N18-1074/.

TTB. Importing bottled alcohol beverages into the united states. Alcohol and Tobacco Tax and Trade Bureau, International Affairs Division, 2025. URL https://www.ttb.gov/import-export/itd/importing-bottled-alcohol-beverages-into-the-united-states. Guidance TTB G: 2015-3; updated November 10, 2025; accessed September 20, 2026.

Tanja Vähämäki, Joona Mäntyvaara, Paavo Nevalainen, and Jukka Heikkonen. B2B opportunity forecasting with explainable segmentation and cost-sensitive prediction. *Applied Intelligence*, 56:425, 2026. doi: 10.1007/s10489-026-07401-z.

Jialong Wu, Wenbiao Yin, Yong Jiang, Zhenglin Wang, Zekun Xi, Runnan Fang, Linhai Zhang, Yulan He, Deyu Zhou, Pengjun Xie, and Fei Huang. Webwalker: Benchmarking LLMs in web traversal. *arXiv preprint arXiv:2501.07572*, 2025.

Migao Wu, Pavel Andreev, and Morad Benyoucef. Profiling before scoring: A two-stage predictive model for B2B lead prioritization. *Journal of Personal Selling & Sales Management*, 46(3):313–329, 2026a. doi: 10.1080/08853134.2026.2633231.

Mingyan Wu, Han Yang, Omer Ben-Porat, and Yftah Ziser. When knowledge is not free: Cost-aware evidence selection in retrieval-augmented generation. *arXiv preprint arXiv:2606.02245*, 2026b. URL https://arxiv.org/abs/2606.02245.

Shuo Yang, Caren Han, Yihao Ding, Shuhe Wang, and Eduard Hovy. ToolTree: Efficient LLM tool planning via dual-feedback Monte Carlo tree search and bidirectional pruning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/3428db291b383208dce69ed572f2b437-Abstract-Conference.html.

Xinyu Zhou, Chang Jin, Carsten Eickhoff, Zhijiang Guo, and Seyed Ali Bahrainian. When silence is golden: Can LLMs learn to abstain in temporal QA and beyond? In *International Conference on Learning Representations*, 2026.

A Complete Task Specification

The qualification label operationalizes a declared weighted score; commercial transaction success is a separate outcome. The six role weights are $(.20, .20, .15, .15, .15, .15)$ for cultural-IP, beauty, and programmable-power domains. Alcohol uses $(.20, .15, .15, .15, .20, .15)$, increasing authorization weight while reducing category weight. The contact and exclusion thresholds are .72 and .35. Scores in between are intrinsically verify, even with all predicates known. Identity is fixed as resolved in the controlled benchmark; identity recovery is therefore outside the measured task.

Table 1's professional sources ground the vocabulary of evidence roles. For example, Medicube's direct-purchase U.S. policy illustrates returns and replacement, while AMETEK's U.S.-market service-contract brochure illustrates calibration and support. The earlier practitioner-reviewed candidate rubric addressed evidence sufficiency, action selection, and responsibility for further verification. Together with the cited professional sources, this design input informs the controlled obligations, which retain a common six-predicate structure across domains. Gold facts and witness spans are generated programmatically for the resulting instances.

There are four content-bearing source bundles: buyer role/market presence (cost 1.15), category/channel (0.90), domain-specific capability (0.75), and fulfillment (0.65). A distractor notice costs 0.35. Bundles whose facts are fully visible initially are removed. Costs are designed comparison units; retrieval charges, latency, and human verification effort require separate measurement. Rich episodes initially expose documents for two rotated facts; the extractor determines which cards are admitted. Sparse episodes expose none. Remaining documents are revealed only by acquisition. GreedyCov retains the identifier `Coverage` in released code and records. S2 retains its original role-aware priors. The hosted public catalog replaces identifiers and source names with opaque bundle labels and replaces all listed predicate priors with $(.4, .4, .2)$.

Generated facts and held-out surfaces. Each domain contains 20 latent profiles, with an assigned language and one of four evaluation surfaces. The rich/sparse pair preserves its six truth values and qualification label, but entity identifiers, initial observations, and the distractor variant can differ. These controlled interventions isolate visibility effects; population-grounded missingness models can provide a transfer evaluation. Canary texts use field notes and short memos; evaluation uses prose records (narratives), compact registries (named fields), public bulletins (announcements), and table-like text (labeled rows). Localized predicate labels prevent the English schema key from serving as an unintended language shortcut. All labels, spans, and visibility rules are inspectable generator outputs.

Witness validation. An admitted card has a predicate, support/contradiction state, source tier and authority flag, an absolute Hypertext Transfer Protocol Secure (HTTPS) source identifier, publication time, a document hash, a short source span, and a verification-rule identifier. Visibility requires a known publication date within the cutoff. Explicit negative evidence requires authoritative status. Hashes establish provenance integrity; claim-state accuracy is scored against gold. Missing or conflicting evidence leaves a predicate unresolved. Hypothetical transitions update unresolved, nonconflicted coordinates; active conflict repair would extend the finite-state acquisition controller.

B Conditional Bound Guarantee and Planning Objective

For each predicate, let $Z_k = \{1\}$ for a correct supporting singleton, $\{0\}$ for a correct contradicting singleton, and $\{0, 1\}$ for absence or conflict. If $z_k \in Z_k$, $\min Z_k \le z_k \le \max Z_k$. Multiplication by nonnegative w_k and summation establish Equation 1. The endpoints are sharp over the rectangular product $\prod_k Z_k$; unmodeled dependence or logical constraints could yield tighter bounds. When $L \ge \theta_+$, every feasible complete score meets contact; when $U < \theta_-$, every feasible complete score meets exclusion. An identity failure blocks both. A consistent additional card refines Z_k to a subset, so its minimum cannot decrease and maximum cannot increase. A contradictory arrival deliberately reopens the coordinate, placing that update outside the refinement premise.

The guarantee is a deterministic implication conditional on valid facts. Probabilistic coverage for language-model extraction would require additional assumptions and calibration. Incorrect singleton evidence may exclude the truth; exact-string matching and authority metadata verify specific

parts of the contract while semantic validity remains an empirical property. Incorrect admission and downstream decision metrics measure distinct failure modes at this boundary.

Why plan beyond coverage? Consider a state whose qualification threshold can only be crossed after two distinct obligations are established. Recovering one low-weight, easy predicate can improve coverage without changing any feasible terminal decision under the remaining budget. The first coordinate of Equation 4 instead scores complete acquisition sequences by the probability of crossing a decision boundary. The second coordinate orders unresolved leaves; the third breaks remaining ties by cost. This lexicographic preference is an explicit design choice specified before evaluation. With at most five finite actions, exact enumeration is practical. Large, dynamically generated catalogs would require a separately evaluated approximation.

The outcome model factorizes per-predicate probabilities and keeps already resolved or conflicted predicates unchanged in hypothetical transitions. Action-outcome probabilities remain fixed throughout an episode. Observed-card merging can still produce conflicts. GreedyCov lacks the same value-based stopping option, so the measured comparison evaluates the complete policies, including stopping. Appendix F isolates planning and stopping for the completed S2 cohort. Matched-stop and prior-sensitivity interventions for the hosted cohort would require additional executions beyond the completed trajectories retained for the present comparison.

The verify penalty in Equation 3 makes the standalone ambiguity score sensitive to unresolved decisions. In the lexicographic recurrence, equal first-coordinate values imply equal expected verify probabilities, so the constant penalty cancels when comparing second coordinates. Resolution is prioritized explicitly by the first coordinate; the matched ablation evaluates that stopping behavior. All comparisons retain the frozen ambiguity penalty and lexicographic ordering.

C Implementation and Study Provenance

Open-weight configuration. The exact checkpoint identifiers and revision prefixes are shown in Table 6; full hashes are retained in the model manifest. All models are frozen, with BF16, greedy decoding, maximum 256 new tokens, batch size one, and seed 20270817. Qwen disables thinking through its frozen chat-template setting. Model-specific loaders and templates are preserved. Qwen's 4B label describes its language component; the repository includes vision weights. Gemma E4B has a dense architecture with 4.5B effective parameters and 8B total parameters including per-layer embeddings.

Table 6: Pinned checkpoints. Citations identify the evaluated model families; revision prefixes identify the exact stored model checkpoint revisions.

Model	Repository and reference	Revision prefix
Qwen3.5-4B	`Qwen/Qwen3.5-4B` (Qwen Team, 2026)	`851bf6e806ef`
Gemma 4 E4B-IT	`google/gemma-4-E4B-it` (Gemma Team, 2026)	`ee0ef6023621`
Phi-4-mini	`microsoft/Phi-4-mini-instruct` (Microsoft et al., 2025)	`09ccd78d4b1b`

The reported S2 cohort is the complete A100 matched-software cohort declared on 11 September. All 48 chunk provenance records (6,720 evaluation outputs) report one NVIDIA A100-SXM4-40GB GPU. Legacy T4 outputs are excluded. A dependency closure was captured per backbone and restored across resumed inference; retained hardware/driver receipts document session-specific environments. Qwen/Gemma use the locked Transformers 5.16.1 stack; Phi retains its model-specific Transformers 4.49.0/Accelerate 1.3.0 compatibility path. Torch 2.11.0 is part of the declared bootstrap. Runtime interruptions and the switch to resident inference preserve the frozen scientific inputs; session interruptions make wall-clock hardware comparisons unsuitable. The selector transfers unchanged from calibration in the earlier A100 environment, maintaining separation from locked-cohort evaluation labels. Controller replays use CPU computation over retained outputs. The hosted Astra service's accelerator is provider-controlled and is separate from the measured A100 cohort.

Calibration and development separation. An earlier development study (S1) motivated heterogeneous interface selection; the subsequent sealed study is denoted S2. S2 canaries contain five

languages, three states, two surfaces, and four replicates: 120 requests per interface/backbone. The 720 canary generations are excluded from evaluation. The sealed selector uses programmatic labels, paired exact tests, fixed multiplicity and effect thresholds. The realized choices are reported in Table 7. The resampling intervals condition on these fixed deployment choices; uncertainty from independently repeating calibration remains outside the estimand of these intervals.

Table 7: Frozen CCIR deployment choices. F: Free; A: Anchor; D: dual agreement. The doubled cost of D is included in counterfactual deployment calls.

Backbone	German	English	French	Japanese	Korean
Qwen	A	A	A	A	A
Gemma	D	D	D	F	D
Phi	A	A	A	F	F

The sealed primary criterion remains unchanged. Full replay was performed after observing S2 and is exploratory. The hosted follow-up was frozen before its own model outputs, but designed after S2 and uses a selected subset of the controlled corpus. An operational record notes limited unintended exposure to a Qwen quality summary before the full analysis; artifact hashes establish integrity while the exposure qualifies the historical blinding. The previously specified endpoints and selection rules remain fixed despite that limited exposure to the quality summary.

Hosted execution. The hosted manifest records Codex CLI `0.155.0-alpha.2.6`, requested model `gpt-6-astra`, effort `ultra`, and the same requested model/effort for any child agents. Isolated inputs disable external search, personal memory, and access to the research corpus; evaluated requests begin with standalone, experiment-specific inputs. The first retained response is parsed without manual semantic repair or best-of-response selection. One documented network recovery preserved the failed attempt and request identity before continuing. Operational transport failure, malformed output, and scientific verify outcomes remain distinct. Completion is 560 shared extraction requests and 560 unique episode–method pairs across 40 independent latent profiles.

D Adapted Agent Implementations

All methods in Table 2 receive the same qualification rules, initial cards, affordable source catalog, and fixed maximum source budget. An acquisition reveals only its own cached cards. Repeated access to a paid bundle costs zero, for every method. A global ledger charges acquisitions even when a search branch is later pruned. For branching methods, the common readout uses the union of globally paid evidence; the native answer uses the selected trajectory's observations. Thus native/common differences can include both readout behavior and branch-level evidence aggregation.

ToolTree adaptation. Following Yang et al. (2026), the controller evaluates available tools before execution, expands promising children, selects using a pre-score-weighted upper-confidence rule, executes a real paid source, then evaluates/prunes and backpropagates mean values. The budgeted implementation uses top-$k = 2$, at most four rollouts and depth three, pre/post thresholds .3/.4, and exploration coefficient 1.0. Each evaluator is a separate hosted request. A final answer uses only the selected executed trajectory, with hypothetical and pruned source contents excluded. These caps differ from the original paper's larger search settings.

Fleet of Agents adaptation. Following Klein et al. (2025), two particles undergo up to three rounds with one mutation per round. The implementation retains enforced movement, sudden-death replacement, cached heuristic values, discounted historical backtracking ($\gamma = .5$), and seeded resampling with replacement (temperature .25). Values are model estimates from observed evidence, never hidden qualification rewards. Source use is globally charged across particles.

ATLAS and MindSearch adaptations. ATLAS retains constraint analysis, planning, verification, and new retrieval for an unresolved information gap (Choi et al., 2026); benchmark proof obligations replace travel constraints. MindSearch retains a graph planner, source-search worker, dependency tracking, and final integration (Chen et al., 2025); finite catalog lookup replaces open-web search,

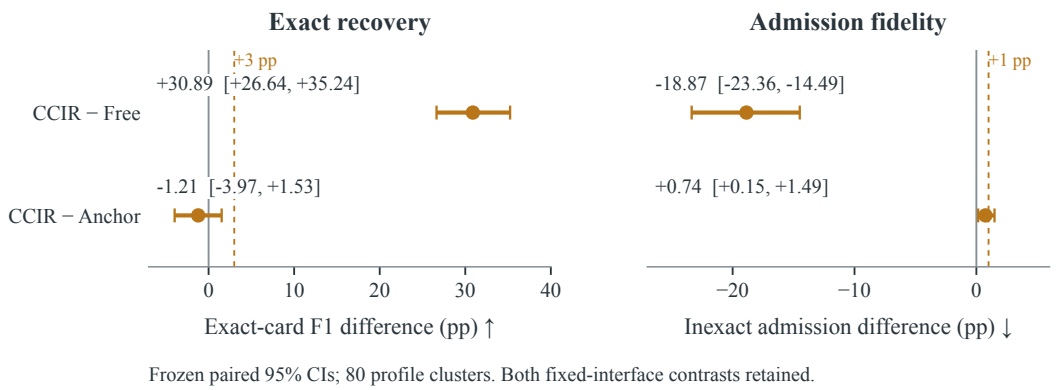


Figure 6: **The two fixed-interface contrasts answer different questions.** CCIR's large gain over Free and its Anchor tradeoff are assessed separately. Intervals are the sealed profile-paired 95% intervals; the benefit point threshold is $+3$ pp and the incorrect-admission upper-bound margin is $+1$ pp for each prespecified paired comparison against a fixed extraction interface.

JavaScript Object Notation (JSON) graph actions replace Python graph construction, and workers execute serially. The exact prompts, role schemas, graph/constraint state, invalid-response handling, and call caps are frozen in the released controller modules. Direct chooses an acquisition or terminal decision without these decompositions. Information isolation holds across methods: hidden successful trajectories, ground-truth judging, and cross-episode label updates are excluded.

The comparison measures adapted mechanisms within the same finite source catalog, complementing the original papers' open-web or broader tool settings. Logical controller calls count observable requests; token budgets and hidden reasoning floating-point operations (FLOPs) can differ. Latency and energy depend on additional execution factors. CPU controllers incur planning cost and use the same shared extracted evidence as the hosted agents.

E Complete Statistical and Metric Results

For exact extraction, a true positive matches the expected state and full immutable gold span. A wrong admitted card adds a false positive and leaves any expected gold card unrecovered. Unresolved outputs are contract-correct only when well formed and no card is expected; parser rejection is incorrect. This separates contract correctness from F1. Precision divides exact true positives by admitted cards, while recall divides them by expected gold cards. An incorrect admitted card can therefore reduce both quantities. State accuracy scores all requests, with rejected outputs counted as incorrect. Output validity includes admitted cards and well-formed unresolved responses, regardless of gold correctness. A policy can improve validity or admitted-card precision by abstaining more often while losing recall. Correct unresolved outputs contribute to state accuracy and contract correctness, while exact-card F1 evaluates recovered cards. Prec., State acc., and Valid denote precision, state accuracy, and accepted-output rate. Inexact divides incorrect admissions by all requests, including state or full-span mismatches; $1 - \mathrm{Inexact}$ is the fraction without such admission, while precision uses admitted cards. Cov. is decision coverage; Sel. acc. is selective accuracy, and Sel. error is its complement among resolved episodes. Confidence intervals use the abbreviation CI. Table 8 preserves both primary contrasts: the substantial Free-relative gain coexists with an unmet sealed joint criterion requiring improvement against both fixed interfaces.

Table 8: Sealed S2 primary contrasts. F1 and incorrect-admission differences are pp; the adjusted p-values test contract correctness; F1 has separate descriptive intervals.

CCIR minus	Δ exact F1 [95% CI]	Holm $p^{\dagger}$	Δ inexact [95% CI]
Free	+30.89 [+26.64, +35.24]	< 0.0001	-18.87 [-23.36, -14.49]
Anchor	-1.21 [-3.97, +1.53]	0.8945	+0.74 [+0.15, +1.49]

Table 9 supplies precision, recall, state accuracy, and output validity for every extraction policy. Precision and recall distinguish informative evidence from well-formed abstention. The oracle is a label-assisted extraction reference with access to test labels; its downstream decisions can still depend on the selected acquisition policy and the resulting controller state.

Table 9: Extraction diagnostics, equally averaged across the fixed three backbones. Rates are percentages; Calls/request excludes calibration. Oracle* is nondeployable.

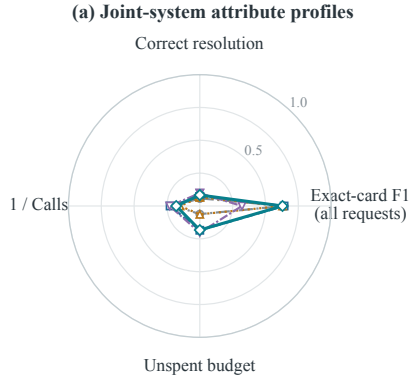
Policy	F1 ↑	Prec.	Recall	State acc.	Valid	Inexact ↓	Calls ↓
Free	32.08	52.92	24.48	51.55	81.55	21.40	1.000
Anchor	64.17	93.95	51.04	53.69	83.72	1.79	1.000
CCIR	62.96	92.62	49.79	54.70	89.32	2.53	1.267
Point	64.54	91.45	51.08	55.83	88.01	3.42	1.000
Validity	58.40	80.73	46.56	57.05	89.43	7.98	1.000
Backbone	59.74	94.16	46.60	50.71	84.73	1.70	1.333
No penalty	61.53	93.91	48.26	53.24	89.70	2.08	1.333
Oracle*	66.92	92.74	54.03	57.59	88.18	2.62	1.000

Selector component controls. Backbone pools all 120 canaries and removes language-specific selection while retaining the paired test, effect threshold, and dual-agreement fallback. No penalty replaces $C - W$ with C while preserving the selection hierarchy. Point selects the sign of the language-specific mean $C - W$ difference, choosing Anchor on ties; this removes significance and effect gates as well as pooled/dual fallback. Validity instead selects the interface with the higher accepted-output rate, also choosing Anchor on ties and replacing the hierarchy. Point and Validity are alternative selector designs rather than isolated single-component removals. A dual-agreement-only removal remains a separate ablation. Table 9 retains every control and metric: components can improve one attribute while trading another. These controls use the frozen outputs and calibration data, keeping backbone generation identical across policies.

Joint-system diagnostics and resource interpretation. Figure 7 crosses Anchor/CCIR extraction with GreedyCov/PIER acquisition on the same S2 profiles and retains Free+PIER as a fixed-interface reference. The combined system improves on CCIR+GreedyCov, while Anchor+PIER uses fewer calls and Free+PIER has higher observed CRDR at lower resource use. This diagnoses composition, with the operating-point tradeoffs retained. Exact-card F1 covers all extraction requests; decision and resource metrics follow the acquired episode trajectory.

Source cost sums bundle charges in designed acquisition units; compute accounting is reported separately. In the hosted study, a deterministic PIER controller makes zero additional model calls yet still acquires documents. It uses less source cost than its common-CRDR ties, GreedyCov and MindSearch*, while several lower-resolution methods spend less. The radars report *unspent budget*, $1 - \bar{C}/2.45$, and inverse extraction calls, $1/\bar{K}$, using mean cost $\bar{C}$ and calls $\bar{K}$. Figure 5(b) uses calls per request; panel (c) uses calls per episode. Each resource axis is interpreted individually. Call accounting excludes calibration and, for hosted controller calls, the shared extraction stage. S2 calls are counterfactual uncached deployment counts: every initial rich-view document and every document in an acquired bundle contributes one call, or two under dual agreement. Both interfaces were generated once; replays reuse their outputs, whereas policy calls describe uncached deployment demand. Initial evidence has zero acquisition charge but still incurs extraction, and bundle prices vary independently of document count. These conventions explain the differing source-cost and call orderings. The hosted radar uses $1/(1 + K)$ for additional controller calls and $1 - C/2.45$ for unspent source budget. Its zero-controller-call condition therefore maps to one while retaining the shared extraction charge in the experimental accounting.

Paired uncertainty. Profile resampling retains both visibility views, all requests, and all compared policies. S2 strata contain 20 profiles per domain; the hosted strata contain 10. Sign-randomization tests assume profile-level paired sign exchangeability within the generated cohort. Bootstrap intervals are nominal percentile intervals conditional on the fixed generators, backbone set, selected interface policy, and retained outputs. Run-to-run hosted variation and future-model uncertainty require repeated system evaluations. The analysis separately records valid and undefined ratio draws.



(b) Downstream outcomes in original units

Extraction + acquisition	CRDR (%) ↑	Cost ↓	Calls ↓	Error (%) ↓
CCIR + PIER	**8.33**	**2.001**	**5.57**	**0.00**
CCIR + GreedyCov	6.67	2.294	6.64	0.00
Anchor + PIER	7.29	1.997	4.42	0.00
Anchor + GreedyCov	5.83	2.299	5.25	0.00
Free + PIER	9.79	1.952	4.36	0.21

Equal-weight mean of three backbones; the same 80 paired profiles per backbone.

Unspent budget = 1 - Cost / 2.45. Calls are extraction calls / episode; resource axes are not utility.

Figure 7: **Combined systems on the same S2 cohort.** Three-backbone macro means; teal identifies CCIR+PIER. Actual decision/resource values accompany the radar. All radial axes use fixed $[0, 1]$ scales: CRDR, exact-card F1, unspent source budget $1 - \bar{C}/2.45$, and inverse extraction calls $1/\bar{K}$. Larger resource coordinates mean less consumption; each coordinate is interpreted separately. These reported resource values exclude the cost of interface calibration.

Zero observed errors can produce a degenerate empirical interval, while population risk remains uncertain beyond the finite set of evaluated profile clusters.

Table 4's subscript standard errors are an additive descriptive summary of the frozen outputs. With seed 20260921, 10,000 shared within-domain profile resamples retain both visibility views and all requests across models and policies. The standard error is the sample standard deviation of bootstrap estimates, conditional on the fixed checkpoints and interface choices. Each macro draw first computes per-backbone metrics, then averages them equally. These errors quantify profile-sampling variation rather than repeated-run variation; this additive summary leaves the original paired intervals, statistical tests, and prespecified success criterion unchanged.

The full exploratory family contains 57 contrasts times three additive endpoints. The sealed S2 correctness tests use 100,000 paired profile sign flips; full-replay and hosted tests use 20,000. The hosted family covers six common-readout CRDR contrasts. The identical adjusted cost values reflect the 20,000-draw Monte Carlo resolution floor and Holm correction; their effect precisions remain distinct. Table 10 includes all 21 deployable acquisition contrasts. Nominal CRDR intervals describe individual contrasts; simultaneous quality-preservation guarantees would require an inferential design that explicitly controls uncertainty across the complete contrast family.

Tables 11–13 report all 48 replay configurations, including the six separated oracle references. Point-Pareto membership is computed on CRDR, erroneous resolution per episode, source cost, and deployment extraction calls. Equal coordinate vectors remain tied, with oracle references excluded from membership. Nondominated policies share a finite observed frontier, which permits multiple tradeoff optima across the four reported decision and resource coordinates.

Hosted primary comparisons. Table 14 reports all six predeclared common-readout contrasts. Several quality differences have positive nominal bootstrap lower bounds; primary conclusions follow the discrete, multiplicity-adjusted tests. These procedures have different inferential constructions. Cost intervals remain secondary descriptions, without additional cost p-values. Native comparisons in Table 5 are descriptive. Under the native readout, Direct and ToolTree each falsely exclude one of 28 gold-contact cases; FoA has no gold-contact false exclusion but one unsafe exclusion among 56 non-exclude cases. PIER correctly resolves 14 cases and excludes no gold-contact case under either readout; all seven common readouts have zero observed incorrect resolutions. These denominators are retained alongside rates to distinguish limited coverage from risk guarantees.

Table 10: All deployable PIER-minus-GreedyCov contrasts. Differences in CRDR are pp; cost uses designed units. Intervals are nominal profile-paired 95% intervals. All cost $p_{\mathrm{Holm}} = .00855$ and all CRDR $p_{\mathrm{Holm}} = 1$ under the 171-test exploratory family.

Backbone	Policy	Δ CRDR [95% CI]	Holm p	Δ cost [95% CI]	Holm p
Qwen3.5-4B	Free	+0.625 [+0.000, +1.875]	1.0000	-0.484 [-0.535, -0.432]	0.0085
	Anchor	+2.500 [-0.625, +6.250]	1.0000	-0.224 [-0.258, -0.193]	0.0085
	CCIR	+2.500 [-0.625, +6.250]	1.0000	-0.224 [-0.258, -0.193]	0.0085
	Point	+1.875 [-1.250, +5.000]	1.0000	-0.237 [-0.272, -0.204]	0.0085
	Validity	+1.875 [-1.250, +5.000]	1.0000	-0.237 [-0.272, -0.204]	0.0085
	Backbone	+2.500 [-0.625, +6.250]	1.0000	-0.224 [-0.258, -0.193]	0.0085
	No penalty	+2.500 [-0.625, +6.250]	1.0000	-0.224 [-0.258, -0.193]	0.0085
Gemma4-E4B	Free	+2.500 [-0.625, +6.250]	1.0000	-0.261 [-0.308, -0.217]	0.0085
	Anchor	+0.625 [-1.250, +3.125]	1.0000	-0.341 [-0.386, -0.297]	0.0085
	CCIR	+1.250 [-1.250, +3.750]	1.0000	-0.358 [-0.411, -0.307]	0.0085
	Point	+1.250 [-1.250, +3.750]	1.0000	-0.309 [-0.356, -0.264]	0.0085
	Validity	+0.625 [-1.250, +3.125]	1.0000	-0.341 [-0.386, -0.297]	0.0085
	Backbone	+0.625 [-1.250, +3.125]	1.0000	-0.401 [-0.451, -0.351]	0.0085
	No penalty	+0.625 [-1.250, +3.125]	1.0000	-0.401 [-0.451, -0.351]	0.0085
Phi4-mini	Free	+1.875 [-1.250, +5.000]	1.0000	-0.267 [-0.309, -0.228]	0.0085
	Anchor	+1.250 [-1.875, +4.375]	1.0000	-0.340 [-0.388, -0.292]	0.0085
	CCIR	+1.250 [-1.875, +4.375]	1.0000	-0.297 [-0.340, -0.254]	0.0085
	Point	+1.250 [-1.875, +4.375]	1.0000	-0.297 [-0.340, -0.254]	0.0085
	Validity	+1.875 [-0.625, +5.000]	1.0000	-0.290 [-0.332, -0.250]	0.0085
	Backbone	+1.250 [-1.875, +4.375]	1.0000	-0.340 [-0.388, -0.292]	0.0085
	No penalty	+1.250 [-1.875, +4.375]	1.0000	-0.297 [-0.340, -0.254]	0.0085

Table 11: Complete Qwen replay. Rates are percentages; deployment calls are per episode. $\mathcal{P}$ marks four-axis point-Pareto membership, using error per all episodes whereas selective error conditions on decisions. Oracle* is a nondeployable reference.

Policy	Router	CRDR ↑	Cov.	Sel. error ↓	Cost ↓	Calls ↓	$\mathcal{P}$
Free	GreedyCov	1.88	1.88	0.00	2.319	5.287	
Free	PIER	2.50	2.50	0.00	1.836	4.225	✓
Anchor	GreedyCov	11.25	11.25	0.00	2.281	5.206	
Anchor	PIER	13.75	13.75	0.00	2.056	4.494	✓
CCIR	GreedyCov	11.25	11.25	0.00	2.281	5.206	
CCIR	PIER	13.75	13.75	0.00	2.056	4.494	✓
Point	GreedyCov	10.00	10.00	0.00	2.285	5.219	
Point	PIER	11.88	11.88	0.00	2.048	4.481	✓
Validity	GreedyCov	10.00	10.00	0.00	2.285	5.219	
Validity	PIER	11.88	11.88	0.00	2.048	4.481	✓
Backbone	GreedyCov	11.25	11.25	0.00	2.281	5.206	
Backbone	PIER	13.75	13.75	0.00	2.056	4.494	✓
No penalty	GreedyCov	11.25	11.25	0.00	2.281	5.206	
No penalty	PIER	13.75	13.75	0.00	2.056	4.494	✓
Oracle*	GreedyCov	11.25	11.25	0.00	2.281	5.206	
Oracle*	PIER	13.75	13.75	0.00	2.056	4.494	

Table 12: Complete Gemma replay, under the same definitions as Table 11.

Policy	Router	CRDR ↑	Cov.	Sel. error ↓	Cost ↓	Calls ↓	$\mathcal{P}$
Free	GreedyCov	11.88	13.12	9.52	2.270	5.181	
Free	PIER	14.37	15.00	4.17	2.009	4.438	✓
Anchor	GreedyCov	3.12	3.12	0.00	2.306	5.269	
Anchor	PIER	3.75	3.75	0.00	1.965	4.388	✓
CCIR	GreedyCov	4.38	4.38	0.00	2.297	9.450	
CCIR	PIER	5.62	5.62	0.00	1.939	7.787	✓
Point	GreedyCov	4.38	4.38	0.00	2.297	5.250	
Point	PIER	5.62	5.62	0.00	1.988	4.419	✓
Validity	GreedyCov	3.12	3.12	0.00	2.306	5.269	
Validity	PIER	3.75	3.75	0.00	1.965	4.388	✓
Backbone	GreedyCov	3.12	3.12	0.00	2.306	10.537	
Backbone	PIER	3.75	3.75	0.00	1.905	8.625	✓
No penalty	GreedyCov	3.12	3.12	0.00	2.306	10.537	
No penalty	PIER	3.75	3.75	0.00	1.905	8.625	✓
Oracle*	GreedyCov	4.38	4.38	0.00	2.297	5.250	
Oracle*	PIER	5.62	5.62	0.00	1.988	4.419	

Table 13: Complete Phi replay, under the same definitions as Table 11.

Policy	Router	CRDR ↑	Cov.	Sel. error ↓	Cost ↓	Calls ↓	$\mathcal{P}$
Free	GreedyCov	10.62	10.62	0.00	2.278	5.200	
Free	PIER	12.50	12.50	0.00	2.011	4.431	✓
Anchor	GreedyCov	3.12	3.12	0.00	2.309	5.269	
Anchor	PIER	4.38	4.38	0.00	1.970	4.375	✓
CCIR	GreedyCov	4.38	4.38	0.00	2.305	5.256	
CCIR	PIER	5.62	5.62	0.00	2.008	4.419	✓
Point	GreedyCov	4.38	4.38	0.00	2.305	5.256	
Point	PIER	5.62	5.62	0.00	2.008	4.419	✓
Validity	GreedyCov	6.25	6.25	0.00	2.298	5.237	
Validity	PIER	8.12	8.12	0.00	2.008	4.425	✓
Backbone	GreedyCov	3.12	3.12	0.00	2.309	5.269	
Backbone	PIER	4.38	4.38	0.00	1.970	4.375	✓
No penalty	GreedyCov	4.38	4.38	0.00	2.305	5.256	
No penalty	PIER	5.62	5.62	0.00	2.008	4.419	✓
Oracle*	GreedyCov	4.38	4.38	0.00	2.305	5.256	
Oracle*	PIER	5.62	5.62	0.00	2.008	4.419	

Table 14: Hosted PIER-minus-baseline common-readout contrasts. CRDR differences are pp; Holm covers six CRDR tests. Cost intervals are secondary and nominal. Each external method is evaluated through its cited task-specific adaptation listed in Table 2.

PIER minus	Δ CRDR [95% CI]	Holm p	Δ source cost [95% CI]
GreedyCov	+0.000 [+0.000, +0.000]	1.0000	-0.123 [-0.151, -0.097]
Direct Astra	+5.000 [+1.250, +10.000]	0.6387	+0.237 [+0.149, +0.324]
ATLAS* (Choi et al., 2026)	+6.250 [+1.250, +12.500]	0.6387	+0.517 [+0.476, +0.559]
MindSearch* (Chen et al., 2025)	+0.000 [+0.000, +0.000]	1.0000	-0.111 [-0.141, -0.078]
ToolTree* (Yang et al., 2026)	+8.750 [+2.500, +15.000]	0.2007	+0.364 [+0.267, +0.468]
FoA* (Klein et al., 2025)	+5.000 [+1.250, +10.000]	0.6387	+0.479 [+0.385, +0.571]

F MATCHED PLANNING AND STOPPING ABLATION

This post-hoc exploratory study isolates two controller components after inspecting S2. The design and implementation were hashed before replay; the sealed S2 success criterion remains unchanged. The 2×2 conditions vary one-step versus full remaining-horizon planning and optional value-based STOP on versus off. Every condition retains the same lexicographic terminal objective, priors, costs, 2.45 budget, identity gate, qualification thresholds, tie-breaking, and frozen CCIR outputs. One-step replans after every acquisition throughout the available budget. STOP off forces acquisition only while unresolved with an affordable action. Resolution, exhausted actions, and budget infeasibility remain mandatory terminals, and the planning-depth boundary retains terminal evaluation.

Full/STOP-on reproduced all 480 historical CCIR+PIER records exactly, including decisions, bounds, action traces, evidence, costs, and logical calls, before the other cells ran. The completed study contains 1,920 replay records, including 1,440 additional controller records, with zero new model calls. All 1,920 records are analyzed within the same 80 paired-profile clusters, each retaining both visibility views and compared conditions.

Table 15: **All component cells with frozen CCIR.** Each row uses 160 episodes / 80 profiles. CRDR, coverage, and erroneous resolution per episode are percentages; Cost and Calls are per-episode means. **Bold**† denotes a lower resource value than STOP off at the same horizon, with Holm $p = .00225$ in the 45-test family. Zero observed error is distinct from a zero-risk guarantee.

Backbone	Horizon	STOP	CRDR	Cov.	Error	Cost	Calls
Qwen	Full	Off	13.75	13.75	0.00	2.277	5.125
Qwen	Full	On	13.75	13.75	0.00	**2.056**†	**4.494**†
Qwen	One	Off	13.12	13.12	0.00	2.254	5.175
Qwen	One	On	13.12	13.12	0.00	**1.834**†	**4.162**†
Gemma	Full	Off	5.62	5.62	0.00	2.233	9.287
Gemma	Full	On	5.62	5.62	0.00	**1.939**†	**7.787**†
Gemma	One	Off	5.62	5.62	0.00	2.224	9.431
Gemma	One	On	5.62	5.62	0.00	**1.638**†	**7.044**†
Phi	Full	Off	5.62	5.62	0.00	2.262	5.138
Phi	Full	On	5.62	5.62	0.00	**2.008**†	**4.419**†
Phi	One	Off	5.62	5.62	0.00	2.264	5.244
Phi	One	On	5.62	5.62	0.00	**1.760**†	**4.081**†

Matched effects and uncertainty. The five fixed contrasts are Full–One under each STOP setting, On–Off under each horizon, and their interaction $(\text{Full/On} - \text{One/On}) - (\text{Full/Off} - \text{One/Off})$. Each contrast is evaluated on CRDR, cost, and calls for each backbone, producing one 45-test Holm family. Domain-stratified paired cluster bootstrap intervals use 10,000 draws; two-sided profile sign-flip tests use 20,000 draws and assume sign exchangeability. Seed 20260920 fixes shared draws across conditions. Intervals describe individual contrasts, and macro means summarize the three fixed models; inference conditions on this backbone set.

Table 15 and Tables 16–18 retain every cell and planned test. At each horizon, STOP preserves every individual observed decision while reducing resources. At full horizon, relative cost reductions are 9.70/13.17/11.20% and call reductions are 12.32/16.15/13.99% for Qwen/Gemma/Phi. The corresponding nominal cost intervals are $[-.245, -.197]$, $[-.327, -.263]$, and $[-.279, -.228]$. One-step STOP yields still larger resource savings; full planning with STOP consumes more cost and calls in all three models ($p_H = .00225$). Its quality gain is one additional correct decision for Qwen and none for Gemma or Phi ($p_H = 1$). These results identify optional stopping as a robust efficiency component within this cohort, while preserving the horizon tradeoff and the distinction between observed decision equality and a population noninferiority guarantee.

Table 16: Complete Qwen component contrasts. CRDR differences are pp; Cost/Calls use original units. Brackets are nominal 95% paired-profile intervals. Holm adjusts all 45 tests jointly. Bold isolates significant resource savings from enabling STOP.

Contrast	Metric	Δ	95% CI	Holm p
Full–One / On	CRDR	+0.625	[+0.000, +1.875]	1.00000
Full–One / On	Cost	+0.222	[+0.146, +0.306]	0.00225
Full–One / On	Calls	+0.331	[+0.231, +0.438]	0.00225
Full–One / Off	CRDR	+0.625	[+0.000, +1.875]	1.00000
Full–One / Off	Cost	+0.023	[+0.007, +0.040]	0.19889
Full–One / Off	Calls	-0.050	[-0.075, -0.025]	0.15769
On–Off / Full	CRDR	+0.000	[+0.000, +0.000]	1.00000
On–Off / Full	Cost	**-0.221**	[-0.245, -0.197]	0.00225
On–Off / Full	Calls	**-0.631**	[-0.700, -0.562]	0.00225
On–Off / One	CRDR	+0.000	[+0.000, +0.000]	1.00000
On–Off / One	Cost	**-0.420**	[-0.493, -0.355]	0.00225
On–Off / One	Calls	**-1.012**	[-1.131, -0.900]	0.00225
Interaction	CRDR	+0.000	[+0.000, +0.000]	1.00000
Interaction	Cost	+0.199	[+0.134, +0.270]	0.00225
Interaction	Calls	+0.381	[+0.269, +0.500]	0.00225

Table 17: Complete Gemma component contrasts, with metric definitions and the joint Holm adjustment across all 45 tests matching those used for Qwen in Table 16.

Contrast	Metric	Δ	95% CI	Holm p
Full–One / On	CRDR	+0.000	[+0.000, +0.000]	1.00000
Full–One / On	Cost	+0.302	[+0.223, +0.384]	0.00225
Full–One / On	Calls	+0.744	[+0.550, +0.944]	0.00225
Full–One / Off	CRDR	+0.000	[+0.000, +0.000]	1.00000
Full–One / Off	Cost	+0.009	[-0.012, +0.032]	1.00000
Full–One / Off	Calls	-0.144	[-0.194, -0.087]	0.01400
On–Off / Full	CRDR	+0.000	[+0.000, +0.000]	1.00000
On–Off / Full	Cost	**-0.294**	[-0.327, -0.263]	0.00225
On–Off / Full	Calls	**-1.500**	[-1.662, -1.337]	0.00225
On–Off / One	CRDR	+0.000	[+0.000, +0.000]	1.00000
On–Off / One	Cost	**-0.586**	[-0.664, -0.513]	0.00225
On–Off / One	Calls	**-2.387**	[-2.650, -2.131]	0.00225
Interaction	CRDR	+0.000	[+0.000, +0.000]	1.00000
Interaction	Cost	+0.292	[+0.223, +0.365]	0.00225
Interaction	Calls	+0.887	[+0.694, +1.087]	0.00225

Table 18: Complete Phi component contrasts, with metric definitions and the joint Holm adjustment across all 45 tests matching those used for Qwen in Table 16.

Contrast	Metric	Δ	95% CI	Holm p
Full–One / On	CRDR	+0.000	[+0.000, +0.000]	1.00000
Full–One / On	Cost	+0.249	[+0.171, +0.332]	0.00225
Full–One / On	Calls	+0.337	[+0.231, +0.450]	0.00225
Full–One / Off	CRDR	+0.000	[+0.000, +0.000]	1.00000
Full–One / Off	Cost	-0.003	[-0.020, +0.016]	1.00000
Full–One / Off	Calls	-0.106	[-0.125, -0.087]	0.00225
On–Off / Full	CRDR	+0.000	[+0.000, +0.000]	1.00000
On–Off / Full	Cost	**-0.253**	[-0.279, -0.227]	0.00225
On–Off / Full	Calls	**-0.719**	[-0.787, -0.650]	0.00225
On–Off / One	CRDR	+0.000	[+0.000, +0.000]	1.00000
On–Off / One	Cost	**-0.505**	[-0.582, -0.433]	0.00225
On–Off / One	Calls	**-1.162**	[-1.281, -1.050]	0.00225
Interaction	CRDR	+0.000	[+0.000, +0.000]	1.00000
Interaction	Cost	+0.251	[+0.182, +0.326]	0.00225
Interaction	Calls	+0.444	[+0.338, +0.556]	0.00225

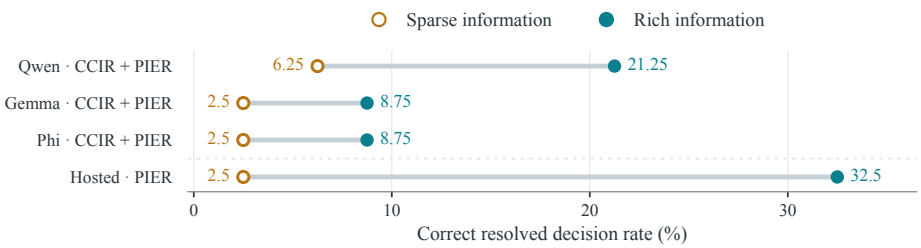


Figure 8: **Paired visibility views expose the evidence-access challenge.** Rich/sparse CRDR (%) for S2 CCIR+PIER (80 profiles) and hosted Anchor+PIER (40 profiles). Pairs preserve latent qualification within each separate cohort; links describe within-cohort contrasts, and causal effects across the two studies remain outside the scope of this comparison.

G Heterogeneity, Prior Work, and Transfer Boundary

Table 19 reports all prespecified subgroup F1 means for the three principal policies. These descriptive cells share profiles and models; the sealed target stays fixed. Figure 8 contrasts paired rich/sparse views. Hosted PIER's 32.5/2.5% CRDR shows that preserving uncertainty and acquiring sufficient evidence address distinct aspects of qualification coverage.

Table 19: S2 subgroup exact-card F1, equally weighted across frozen backbones. All levels of each prespecified grouping are shown; these are descriptive diagnostics.

Grouping	Level	Free F1	Anchor F1	CCIR F1
Language	German	36.98	77.34	70.67
	English	18.75	80.20	73.45
	French	24.93	63.93	60.53
	Japanese	59.50	61.38	72.09
	Korean	5.86	20.64	25.47
Domain	Alcohol	32.39	64.36	62.98
	Beauty	28.71	61.12	60.99
	Cultural-IP	31.72	62.12	61.13
	Programmable power	35.20	68.61	66.40
Surface	Compact registry	41.34	65.86	68.80
	Prose record	47.85	73.98	75.84
	Public bulletin	0.00	51.06	34.20
	Table-like	37.03	63.08	64.50
Visibility	Rich	32.06	64.22	62.59
	Sparse	32.10	64.12	63.33

Relation to the earlier position paper. Nam et al. (2026) frames decision-grade curation across four export domains through provenance and time validity, uncertainty-aware qualification, unknown-not-false hard gates, and budgeted decision value. The present work adds explicit interval states and conditional guarantees, an executable lexicographic router, programmatic interface contracts and a hierarchical selector, completed frozen open-weight evaluation, and shared-cache hosted comparisons. Earlier practitioner feedback informs the task specification; the present mechanisms and experiments operationalize a controlled subset of the broader agenda. Representative public-document evaluation is the next transfer step.

Scope for transfer evaluation. Public-document testing, conversational rule reasoning (Saeidi et al., 2018), hosted CCIR, broader agent budgets, and H100 replication remain prospective. Completed controlled studies supply the reported results; earlier audits serve development. Country demand, commodity prices, buyer intent, outreach, and longitudinal commercial outcomes lie beyond the implemented task. The release supports reproducible comparisons under controlled qualification and visibility conditions; transfer to commercial populations is the next empirical step.