# CLESO: CRITERION-GUIDED LAMINAR EVIDENCE SEQUENCE ORGANIZATION FOR SPECIFIC DOMAIN BUDGETED PROBATIVE EVIDENCE SELECTION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Evidence selection in domain-specific question answering aims to identify a compact set of text units from a large-scale corpus that jointly supports an answer. Building on high-recall candidate sequences produced by relevance ranking, existing systems improve evidence selection through instructions, iterative reasoning, or set-level objectives. However, evidence set selection in domain-specific QA must locate support that may be dispersed across long candidate sequences and assess whether a candidate set jointly enables a reliable answer without mixing incompatible entity conditions or scopes. Repeatedly assessing alternative sets with an LLM is costly. We propose CLESO, a structural sequence distillation framework that learns laminar candidate sequences to prioritize access to potential support under a limited observation budget before downstream joint selection. During supervision, ontology-grounded criteria provide signal for candidate qualifications and cross-evidence dependencies, and a Laminar compiler turns these signals into local ordering, head-admission, and boundary-separation targets. At inference, a lightweight Transformer refiner predicts residual scores from inference-available features, and a fixed decoder recovers intervals from the resulting sequence. The recovered intervals guide budgeted local observation before LLM-based package selection, allowing the learned structure to guide evidence access without executing criterion search at inference On BioASQ, QASPER, and LegalBench-RAG, CLESO-CP gains an average of 1.45 evidence-F1 points over the strongest baseline, with a 3.25k-token observation cap on queues averaging 1M tokens.

## 1 INTRODUCTION

Question answering in specific domains relies on external evidence grounding for factual accuracy and verifiability (Lewis et al., 2020; Gao et al., 2023; Fan et al., 2024), making it a fundamental capability of domain-specific language systems. To access evidence in a large corpus, retrieval models and systems organize text units into a high-recall relevance candidate sequence based on the question (Karpukhin et al., 2020; Khattab & Zaharia, 2020; Jin et al., 2023). These sequences allow existing systems to improve evidence selection through instruction-conditioned retrieval (Asai et al., 2023; Weller et al., 2025), iterative reasoning (Trivedi et al., 2023; Li et al., 2026), and set-level objectives (Lee et al., 2025; Wu et al., 2026b). However, the selected sources must jointly support the answer under compatible conditions and scopes. Alternative evidence sets may satisfy the same information requirements, but their adequacy depends on whether they provide complete support under compatible conditions and scopes. Finding such a set within a long candidate sequence can require extensive reading and repeated evaluation of evidence combinations, even with long-context LLMs. Therefore, the challenge is to identify sufficient support within an observation budget while controlling the computational cost of evidence discovery.

Fig. 1 shows the bottlenecks of probative evidence selection in the long candidate evidence sequence. Annotated targets remain dispersed (Fig. 1(A) left), so early prefixes can miss support; yet answer success from full prefixes does not reliably improve as more text is read (Fig. 1(A) right). The target-only diagnostic separates the availability of annotated evidence from the ability to answer with the surrounding text. Processing more candidates also has a cost; in Fig. 1(B), methods with similar evidence F1 differ markedly in runtime and input-token use. The LegalBench-RAG case in Fig. 1(C) shows this semantic difficulty: the top-ranked passage is relevant but misses the contractual excep-



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.

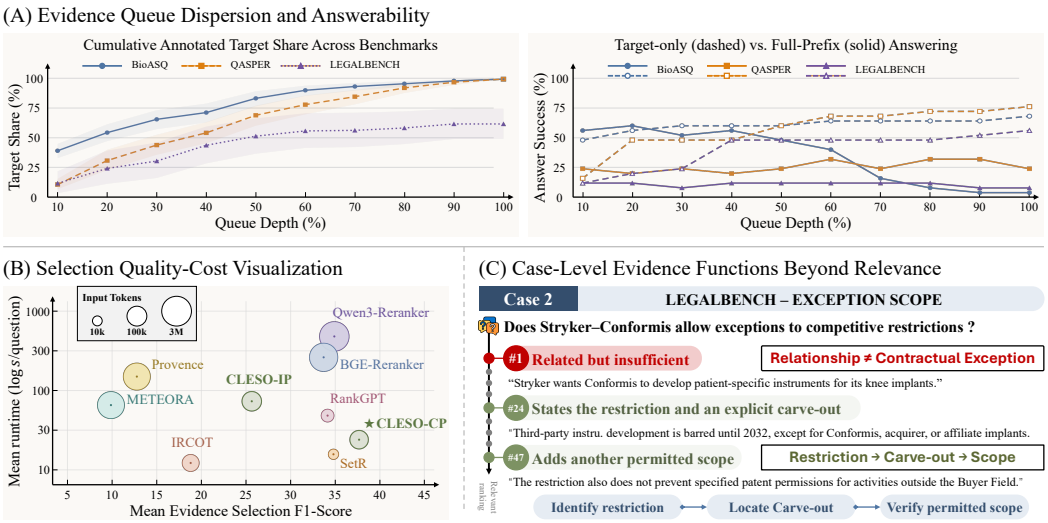


Figure 1: **Evidence selection bottlenecks.** (A) Annotated-target recall by normalized queue depth (left) and answer success using label-visible target-only evidence (dashed) or full prefixes (solid; right). Left shading shows confidence intervals; context-limit failures count as incorrect. LegalBench-RAG queues are capped at 1M tokens, with recall measured against all annotated targets. (B) Bubbles encode equal-weight three-benchmark means of evidence F1 (x), method-native serial-equivalent runtime (y), and input tokens per question (log-scaled area). (C) Arrows map ranked LegalBench-RAG passages to the restriction, its carve-out, and permitted scope.

tion, while lower-ranked passages identify the carve-out and its scope. These observations motivate a learnable access structure that exposes dispersed support within a fixed observation budget and preserves evidence roles for probative set selection without reading the full queue. [1]

The first challenge is the *joint support evidence identifiability* in a high-recall sequence. Fig. 1(A) shows that domain-specific questions require passages with distinct candidates; retaining enough candidates for recall can leave these passages far apart among topical distractors. Ontology and schema attributes provide reusable cues, but their overlap across roles varies by case, which cannot establish whether passages jointly satisfy the question under compatible conditions. Supervised rerankers reorder candidates by learned relevance (Zhang et al., 2025), while evidentiality signals and LLM feedback add support cues (Song et al., 2024; Wu et al., 2026b). PACE frontloads complementary information through semantic coverage (Cai & Zafarani, 2026), and BESC constructs cross-library evidence sets through aspect coverage control (Li, 2026). Yet, the remaining problem is organizing the dispersed, potentially compatible joint assessment on the long sequence.

The second challenge is *budgeted cost* access to a stable support evidence set. A single global prefix with a score from compact features may miss support in separated regions, whereas reading the full queue and repeatedly assessing candidate sets can be costly. SetR (Lee et al., 2025) distills set-level decisions but still operates on supplied passages; context compression processes retrieved text before shortening the answer context (Xu et al., 2024), while iterative methods add retrieval and reasoning steps (Trivedi et al., 2023; Li et al., 2026). The resulting access problem is to distribute a bounded source-text budget across promising regions before joint evidence selection.

To address these challenges, we propose CLESO, a structural sequence distillation framework for observation-budgeted probative evidence selection. During supervision, forward ontology-based criterion search compiles feasible intervals with exact criterion–target agreement on the target runs they cover, and uses admissible relevance-score gaps to locate their boundaries. A lightweight Transformer then learns residual scores for local priorities and boundary separation. At inference, the recovered intervals guide probe observation and proposal-conditioned package selection. Criterion search is confined to supervision, while inference limits the candidate source text passed to joint selection to the observation pool. We also characterize conditions for structural recovery and budgeted witness exposure. On the full BioASQ, QASPER, and LegalBench-RAG Test splits, CLESO-CP exceeds the strongest evidence-F1 baseline by 0.55, 1.93, and 1.86 percentage points, respectively

[1] Additional cases and analysis are provided in Appendix A.

(1.45 points on average). It also leads reference-conditioned PB on all three benchmarks and Answer on BioASQ and QASPER. These results use a 3,250-token observation cap on candidate queues averaging approximately one million source tokens.

- CLESO builds the supervision of local support priorities and interval boundaries from label-visible ontology criteria and admissible relevance-score gaps.
- CLESO distills this structure into a lightweight refiner whose inference scores guide interval recovery, probe-based local observation, and proposal-conditioned package selection.
- We characterize conditions for structural recovery and budgeted witness exposure, then evaluate evidence selection across BioASQ, QASPER, and LegalBench-RAG under fixed token budgets.

## 2 Related Work

**Evidence-aware retrieval and set selection.** Evidence-aware retrieval distinguishes support from distractors (Song et al., 2024), while complementary-evidence models identify distinct answer aspects (Mou et al., 2021). Deterministic gating assesses individual evidence units (Unda, 2026). SetR models joint information requirements (Lee et al., 2025); BESC and QUBO-based selection model aspect coverage, redundancy, and passage interactions (Li, 2026; Singh & Vadlamani, 2026). METEORA uses retrieval rationales (Saxena et al., 2025), Rubric4Setwise uses explicit rubrics with reference-answer assistance in its reported experiments (Jiang et al., 2026), and BAR-RAG trains selection using generator feedback (Sun et al., 2026). These methods assess or select available evidence; CLESO learns which regions of a long queue to expose before set-level selection.

**Structured retrieval and ranking distillation.** RAPTOR, GraphRAG, and StructRAG organize retrieval through hierarchies, entity graphs, and question-adaptive representations (Sarthi et al., 2024; Edge et al., 2024; Li et al., 2025); structured domain knowledge also supports condition-aware retrieval (Lee et al., 2026). RankDistil preserves top ranks through distillation (Reddi et al., 2021), while RankGPT obtains LLM-generated rankings (Sun et al., 2023). CLESO instead distills criterion-derived local priorities and interval separation along an existing ranked sequence.

**Budgeted context construction and adaptive retrieval.** RECOMP compresses retrieved context (Xu et al., 2024), and Provence prunes and reranks it (Chirkova et al., 2025). AdaGReS selects context under a token budget using relevance and redundancy (Peng et al., 2025), while PACE frontloads evidence and adapts reranking effort (Cai & Zafarani, 2026). Cost-aware RAG budgets source-access tiers (Wu et al., 2026a); AB-RAG controls further retrieval (Kamthan, 2026); and IR-CoT and S2G-RAG link retrieval to reasoning or evidence gaps (Trivedi et al., 2023; Li et al., 2026). CLESO allocates an observation budget across learned interval heads before pack-level selection.

## 3 Preliminary

Given an evidence-grounded question $q$ and a corpus $\mathcal{P}$, an upstream retriever returns a candidate relevance sequence $P = (p_1, \ldots, p_n)$, where $p_i \in \mathcal{P}$ denotes a retrieved unit, such as a document or passage. The associated relevance scores $\mathbf{s} = (s_1, \ldots, s_n)$ satisfy $s_1 \geq \cdots \geq s_n$, with original order $\pi_P = (1, \ldots, n)$. We regard $\mathbf{s}$ and $\pi_P$ as part of $P$. A question-answering case is $x = (q, P) \in \mathcal{X}$, where $\mathcal{X}$ denotes the case space. Each evidence span retains its source identity and character offsets. We define the candidate-grounded span universe $\mathcal{S}(P)$ and evidence-set space $\mathcal{E}_x$ as:

$$\mathcal{S}(P) = \{e \mid \exists\, p \in P,\ e \subseteq_{\text{char}} p\}\,, \quad \mathcal{E}_x = \{E \subseteq \mathcal{S}(P) \mid |E| < \infty\}\,.$$

Then, the corresponding supervision is $y = (E^*, a^*)$, where $E^* = \{e_1^*, e_2^*, \ldots, e_m^*\} \in \mathcal{E}_x$ is the marked target evidence, and $a^*$ is the exact reference answer based on $E^*$.

**Probative Evidence Set.** We consider a target evidence set to be *probative* if its source statements jointly provide correct and sufficient support for the answer requirements. Formally, we define the probative evidence-set family $\mathcal{U}_{x,y}$ based on a fixed answerer model $L$ as:

$$\mathcal{U}_{x,y} = \{E \in \mathcal{E}_x \,|\, \mathrm{Rel}_{x,y}(E) = 1\}\,, \quad \mathrm{Rel}_{x,y}(E) = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbf{1}\left[L(q, E; \omega_t) \equiv a^*\right], \qquad (1)$$

where $\mathrm{Rel}_{x,y}(E)$ is the ideal probative condition, which measures the stable answer correctness of the fixed answerer $L$ under stochastic conditions $\omega_t$. Under the idealized witness assumption, $E^*$ is one member of $\mathcal{U}_{x,y}$; successful selection may return any member of this family.

**Selection objective.** We study evidence selection under a source-text observation budget $B_{\text{obs}}$. An observation policy $g$ produces an ordered collection $O = g(x)$ of complete, source-grounded candidate occurrences from $P$, and a selection policy $f$ returns $\hat{E} = f(q, O) \in \mathcal{E}_x$. Both policies use inference-available inputs, with reference evidence and answers restricted to training and evaluation. The objective is to maximize the probability of selecting a probative set from the observed content:

$$(g^*_{B_{\text{obs}}}, f^*_{B_{\text{obs}}}) \in \underset{\substack{g,f:\ \ell(g(x))\le B_{\text{obs}} \\ f(q,g(x))\preceq g(x),\ \forall x\in\mathcal{X}}}{\arg\max} \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\mathbf{1}\left\{f(q, g(x)) \in \mathcal{U}_{x,y}\right\}\right], \tag{2}$$

where $\ell(O)$ sums the source-text token costs of admitted occurrences, and $\hat{E} \preceq O$ requires every selected span to be grounded in the observed source text. The budget $B_{\text{obs}}$ constrains the pool available to downstream selection, while feature construction, independent probing, and repeated model processing are accounted for separately in the cost evaluation. [2]

# 4 CLESO: Learning Laminar Evidence Sequences

To realize budgeted evidence selection in long domain sequences, we propose CLESO, a structural distillation framework. CLESO distills ontology-grounded criterion structure into a sequence access policy. Training-time search supplies interval roles and boundary constraints for a lightweight score refiner. At inference, its scores yield Probe-budgeted local heads, from which proposal-conditioned packing and LLM selection construct the final evidence set. Fig. 2 summarizes the framework.

## 4.1 Ontology-Guided Criterion Search

For each training case $x = (q, P)$, CLESO derives structural supervision from domain concepts, relations, and applicability conditions. An ontology incidence matrix $M_x$ records question-grounded predicates for each candidate, retaining source and entity bindings for relational evaluation. Its inference-time construction uses $q$, $P$, and fixed ontology resources and extraction procedures, without reference evidence, reference answers, or fitted criterion matches. We search a prescribed family $\mathfrak{C}$ of AND–OR–VETO criteria over these predicates, using candidate-level target labels induced by source-span overlap with $E^*$. Let $\eta$ bind criterion roles to candidate identities, and let $A_b$ and $V_b$ contain the required and veto predicates of branch $b$. The criterion is evaluated as:

$$R(\eta) = \bigvee\nolimits_b \left[\left(\bigwedge\nolimits_{\alpha\in A_b} \alpha(\eta)\right) \wedge \neg \left(\bigvee\nolimits_{\alpha\in V_b} \alpha(\eta)\right)\right], \tag{3}$$

where $\bigwedge$ conjunction enforces joint requirements, $\bigvee$ branches admit different support patterns, and veto $\neg$ predicates exclude incompatible matches. Criterion matches supervise access to annotated support, providing a structural surrogate for the ideal probative selection objective in Eq. 2.

To localize the criterion search on $P$, we organize its candidate-index domains as a laminar family.

**Definition 1** (Laminar family)**.** *A family $\mathcal{I}$ of nonempty candidate-index sets is* laminar *if any two members are disjoint or nested: $\forall\ I, J \in \mathcal{I}$, either $I \cap J = \varnothing$, $I \subseteq J$, or $J \subseteq I$.*

The intervals later serve as units for probing and budgeted local reading. CLESO uses the disjoint case of Definition 1, constructing an ordered partition $\mathcal{I}^{\text{cs}}_x$ of all candidate indices into nonempty intervals contiguous in the original sequence. Within each interval, candidate identities form the vertices of a local support hypergraph, and accepted joint criterion matches form its hyperedges; qualification and veto marks specify ordering priorities. The partition is a structural approximation: support relations may still connect candidates from different intervals. Fig. 2(A) shows the process.

Starting from the first unprocessed target, CLESO searches forward for a feasible region. Once such a region $I$ is found, it is extended through the next target position to form $I^+$. The extension is accepted when a criterion matches exactly the targets in that region:

$$I \leftarrow I^+ \quad \text{if} \quad \exists\, R \in \mathfrak{C} :\ \mathrm{Match}(R, I^+) = \mathrm{Target}(I^+), \tag{4}$$

Here $\mathrm{Match}$ and $\mathrm{Target}$ return accepted and annotated candidates in $I^+$; equality gives unit regional precision and recall. When extension fails, CLESO keeps the last feasible region, admits adjacent non-targets that preserve its criterion, and selects the largest normalized original-score gap among admissible boundary positions. Infeasible targets are recorded as uncovered and receive no head supervision; all candidates remain in the final partition. The resulting partition and hypergraph marks are constructed offline for score supervision, so criterion search is not repeated during inference.

[2] Detailed preliminary definitions are provided in Appendix B.

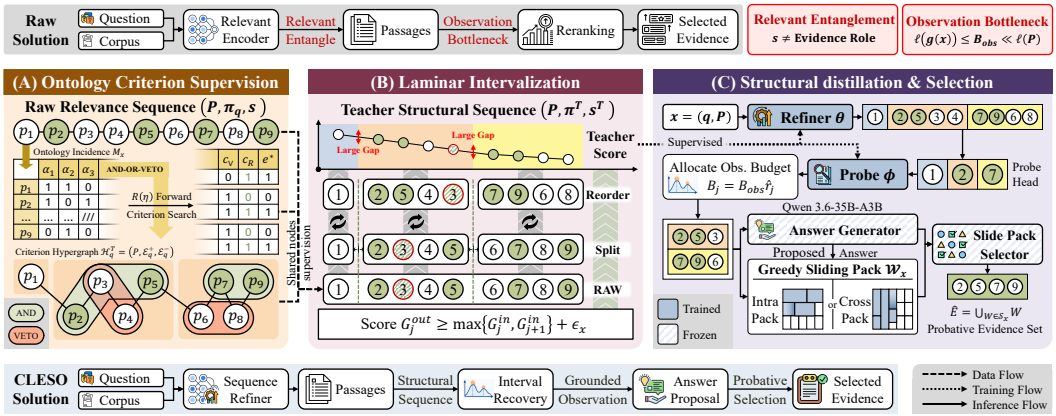


Figure 2: **CLESO overview.** (A) Ontology-grounded criteria identify candidate qualifications and cross-evidence dependencies in the raw relevance sequence. (B) A Laminar teacher compiles these signals into ordering, head-admission, and score-gap separation targets. (C) Learned scores guide interval recovery, probing, and budgeted observation, followed by proposal-conditioned greedy packing and whole-package selection. Candidate layouts are schematic.

### 4.2 Laminar Interval Supervision

Within each interval, criterion-covered targets are promoted to the head, non-target candidates satisfying positive predicates but rejected by VETO are placed at the tail, and all remaining candidates occupy the middle. VETO changes priority, not candidate membership. This ordering concentrates criterion-supported candidates in local prefixes, reducing the prefix depth needed to reach them.

To make these access regions recoverable from scalar predictions, we assign reference scores $s_i^{cs}$ that realize the local priorities within nonoverlapping score bands ordered by the original interval order (Fig. 2(B)). For adjacent intervals $I_j, I_{j+1} \in \mathcal{I}_x^{cs}$, the reference scores satisfy:

$$G_j^{\text{out}} \geq \max\{G_j^{\text{in}}, G_{j+1}^{\text{in}}\} + \epsilon_x, \quad \epsilon_x > 0. \tag{5}$$

where $G_j^{\text{out}}$ is the minimum reference score in $I_j$ minus the maximum in $I_{j+1}$, and $G_j^{\text{in}}$ is the largest adjacent score drop within the locally sorted $I_j$, taking zero for a singleton. This condition separates boundaries from score variation with neighboring intervals, providing the gap signal for inference.

A lightweight Transformer (Vaswani et al., 2017) refiner learns a candidate-level residual scoring function from inference-visible features, with local priorities and boundary separation supplied as training constraints. For candidate $p_i$, the feature vector $\mathbf{u}_i$ concatenates frozen question and candidate embeddings, ontology semantics and extraction states, normalized original ranks, and bidirectional original score differences with validity masks. Projected candidate tokens are contextualized by local self-attention over original neighborhoods and mapped by a shared scalar output head:

$$\delta_\theta(x) = \text{Head}_\theta\Big(\text{Transformer}_\theta^{\text{local}}\big((\text{Proj}_\theta(\mathbf{u}_i))_{i=1}^n\big)\Big) \in \mathbb{R}^n, \quad \mathbf{u}_i = \text{Feat}(q, p_i, M_x, \mathbf{s}, \pi_P),$$

where $\delta_\theta(x)_i$ is the normalized residual for candidate $p_i$. The refined score is $s_i^\theta = s_i + a_x \delta_\theta(x)_i$, where $a_x > 0$ is the floored original score interquartile range used to normalize score-difference features and decoder gaps. Criterion-derived roles and reference scores supervise the refiner through role-ordering and gap penalties, sparse reference-score fitting, and background-drift regularization. At inference, candidates are stably sorted by decreasing $s_i^\theta$, with original position breaking ties. A frozen decoder selects the largest normalized adjacent score gap among admissible cut positions under fixed candidate-count constraints, which also govern when splitting stops. Denote the resulting ordered partition of the refined sequence by $\widehat{\mathcal{I}}_x = (I_1, \ldots, I_{K_x})$. These intervals are recovered from predicted scores without accessing $\mathcal{I}_x^{\text{cs}}$ or target labels.

### 4.3 Interval Head Observation and Selection

Recovered interval heads provide access points at multiple sequence depths, allowing downstream selection to receive local prefixes instead of one long global prefix (Fig. 2(C)). Let $O_j(b)$ return the longest prefix of complete candidate occurrences in recovered interval $I_j$ whose source-text token

cost does not exceed $b$. We probe these heads using equal shares of a separate budget $B_{\text{probe}}$, and a Probe $\phi$ predicts normalized allocation shares $\hat{r}_j$:

$$(\hat{r}_j)_{j=1}^{K_x} = \phi\Big(q,\, (O_j(B_{\text{probe}}/K_x)\,, \text{Stats}(I_j))_{j=1}^{K_x}\Big), \quad \hat{r}_j \geq 0, \quad \textstyle\sum_{j=1}^{K_x} \hat{r}_j = 1, \tag{6}$$

where $\text{Stats}(I_j)$ contains the candidate count, token cost, and refined-score summaries of the interval. The Probe is trained independently using each interval's share of target occurrences, identified by source-span overlap with $E^*$, and remains frozen during refiner training and inference. Its predictions distribute the fixed observation budget as follows:

$$B_j = B_{\text{obs}}\hat{r}_j, \quad O = g(x) = O_1(B_1)\| \cdots \|O_{K_x}(B_{K_x}), \tag{7}$$

where $\|$ denotes ordered concatenation. Prefix costs use the same source-text token accounting as $\ell(O)$, giving $\ell(O) \leq B_{\text{obs}}$. Unused quotas are not redistributed, and probed content enters $O$ only when admitted by the observation quotas. The observed candidates retain their interval and within-interval order. Probative selection requires joint assessment of the observed content, including support distributed across intervals. The fixed answerer generates a proposal $\tilde{a} = L(q, O)$, which conditions greedy construction of a package family $\mathcal{W}_x$. CLESO-IP grows packages within each observed interval in local order, while CLESO-CP traverses observed intervals at each local depth before advancing to the next depth. A frozen Qwen3 reranker (Zhang et al., 2025) assigns each growing package $W$ a proposal-conditioned proxy score from $(q, \tilde{a}, W)$ using the token logit difference. A package is finalized when an appended occurrence lowers its proxy score beyond the fixed tolerance or exceeds a package cap; that occurrence starts the next package. Each variant retains its final partial packages, assigning every observed occurrence to exactly one nonempty package.

A frozen LLM selector receives $q$, $\tilde{a}$, and the source text of all packages in one set-level request. It selects $\mathcal{W}_x^{\text{sel}} \subseteq \mathcal{W}_x$ by jointly assessing support, compatibility, and redundancy, allowing the evidence to verify, complete, or correct the proposal. Selected packages are retained in full; in the following union, each $W$ denotes the source-grounded spans carried by that package:

$$\hat{E} = f(q, O) = \textstyle\bigcup_{W \in \mathcal{W}_x^{\text{sel}}} W \preceq O. \tag{8}$$

The selected source spans are restored to their order in $O$, and $L$ generates the final answer from $(q, \hat{E})$ alone. Therefore, CLESO amortizes criterion search into learned local access and bounds the source text supplied to downstream selection by $B_{\text{obs}}$. Total online cost includes full-sequence refinement, sorting and interval recovery, independent probing, proposal generation, repeated package scoring, selection, and answering, which are evaluated separately from the observation-pool budget.

## 5 Structural Recoverability and Budgeted Access

**Structural transfer and support access.** Let $E^\dagger \in \mathcal{U}_{x,y}$ be a fixed nonempty probative witness, and $A^\dagger$ be a fixed set of candidate indices whose source text covers this witness. For each criterion-derived interval $I_j$, let $h_j$ be the shortest $\mathbf{s}^{\text{cs}}$-ordered prefix cost covering $A^\dagger \cap I_j$ (zero if empty), and set $H = \sum_j h_j$. Let $D_j^\theta$ sum the token costs of candidates outside that prefix that precede any required candidate in the refined local order; the required local prefix then costs at most $h_j + D_j^\theta$. Finally, $b_x$ is the minimum normalized separation between adjacent reference intervals, and $\gamma_x$ is the minimum winning-gap margin along reference decoding; empty minima are $+\infty$.

**Theorem 2** (Structural transfer and budgeted witness access)**.** *Assume the fixed maximum-gap decoder has count-determined legal cuts, traversal, and stopping, and recovers $\mathcal{I}_x^{\text{cs}}$ from $\mathbf{s}^{\text{cs}}$. $\widehat{\mathcal{I}}_x = \mathcal{I}_x^{\text{cs}}$ if $\|\mathbf{s}^\theta - \mathbf{s}^{\text{cs}}\|_\infty / a_x \leq \nu < \min\{b_x/2, \gamma_x/4\}$. For $\varepsilon_{\text{ord}} \geq 0$, $0 \leq \varepsilon_{\text{alloc}} < 1$, and $\forall\ h_j > 0$, the minimum real-valued observation budget $B_{\text{req}}$ needed to include the fixed cover satisfies*

$$B_{\text{req}} \leq \frac{1 + \varepsilon_{\text{ord}}}{1 - \varepsilon_{\text{alloc}}} H, \quad \textit{where} \quad D_j^\theta \leq \varepsilon_{\text{ord}} h_j, \quad \textit{and} \quad \hat{r}_j \geq (1 - \varepsilon_{\text{alloc}}) \frac{h_j}{H}. \tag{9}$$

*Any $B_{\text{obs}}$ at least this upper bound gives $E^\dagger \preceq O$.* [3]

The recovery condition preserves score-band membership and each decoder decision. Within recovered intervals, the required prefix costs at most $h_j + D_j^\theta$; dividing by its allocation share gives

[3] Proofs and cost-accounting details are in Appendix D.

the budget bound. The bound links boundary recoverability and local ordering to the observation budget. Because the Probe learns target-count shares, its token-demand condition must be checked separately. The guarantee establishes a probative witness within $O$; whole-package construction and LLM selection determine whether the final output realizes that support.

**Local reading and computational cost.** Under the same reference order and fixed cover, let $C_j$ be the total source-token cost of $I_j$, and let $j_*$ be the last interval containing required evidence. Reaching $I_{j_*}$ requires a prefix to read the earlier interval. The shortest global-prefix cost obeys

$$B^{\min}_{\text{glob}} - H = \textstyle\sum_{j<j_*}(C_j - h_j) \geq 0. \tag{10}$$

Local access saves the unused suffixes of earlier intervals. The learned policy needs a smaller budget whenever the bound in Eq. 9 is below $B^{\min}_{\text{glob}}$. With one interval, local and global prefix access coincide in the same order. Criterion search is amortized offline, while proposal generation, package selection, and final answering each receive at most $B_{\text{obs}}$ source tokens. Total online cost also includes full-sequence feature construction and refinement, sorting, decoding, independent probing, and repeated package scoring. Structural access therefore controls downstream source-text exposure, with end-to-end efficiency evaluated through complete stage-level accounting.

## 6 Experiments

We organize our evaluation around three research questions:

- **RQ1: Effectiveness.** At a matched final-evidence token cap, how does CLESO compare in evidence selection, answer quality, and reference-conditioned support?
- **RQ2: Budget sensitivity and recorded cost.** How does final-evidence F1 change as the observation budget varies, and what method-native selection cost is recorded on the full Test splits?
- **RQ3: Structural transfer and ablation.** How does the student guide budgeted access, and which observation and packing choices affect final evidence selection?

### 6.1 Experimental Settings

**Benchmarks.** BioASQ (Tsatsaronis et al., 2015) grounds biomedical questions in PubMed snippets, QASPER (Dasigi et al., 2021) links questions to supporting paragraphs in scientific papers, and LegalBench-RAG (Pipitone & Alami, 2024) targets precise spans in legal documents. Their source-grounded annotations permit evaluation of selected support rather than relevance ranking alone. Mean full-Test candidate sequences contain 885,449, 4,904, and 2,088,524 source tokens, respectively, exposing the method to markedly different access regimes under the same observation cap. Table 1 summarizes the data splits of three benchmarks.

Table 1: Details of Benchmark sizes.

| Benchmark | Train | Evaluation | Test |
|---|---|---|---|
| BioASQ | 2,439 | 264 | 260 |
| QASPER | 2,593 | 1,005 | 1,451 |
| LegalBench-RAG | 4,941 | 536 | 1,412 |

**Baselines.** We compare retrieval and reranking models, budget-aware selection and compression methods, and LLM-based ranking and retrieval approaches. Specifically, the model group includes BM25, BGE-Reranker, MedCPT (Jin et al., 2023), and Qwen3-Embedding and Reranker (Zhang et al., 2025), with MedCPT evaluated on BioASQ. The methods include MMR, Adaptive-k, RECOMP-extractive (Xu et al., 2024), Provence-pretrained (Chirkova et al., 2025), AdaGReS (Peng et al., 2025), and fixed-budget PACE-Frontloading (Cai & Zafarani, 2026). The LLM-based approaches includes RankGPT-Qwen (Sun et al., 2023), prompt-only SetR (Lee et al., 2025), METEORA without DPO (Saxena et al., 2025), and IRCoT-Qwen (Trivedi et al., 2023).

**Evaluation Metrics.** Evidence precision, recall, and F1 measure the final selected evidence using each benchmark's frozen source mapping: source-character alignment for BioASQ and LegalBench-RAG, and paragraph-evidence mapping for QASPER. QASPER Answer is official answer-token F1; BioASQ and LegalBench-RAG Answer use frozen model-assisted correctness judgments. Tail Recall measures selected target evidence beyond the same-budget RAW-prefix reference. PB is a reference-conditioned model judgment of whether the selected evidence supports the answer requirements under compatible conditions; it does not certify the ideal probative condition. [4]

[4]Dataset and model details, additional settings, and experiments are provided in Appendix E.

Table 2: Experiment results on BioASQ, QASPER, and LegalBench-RAG at a 3,250-token budget*.

| Method | BioASQ | | | | | | QASPER | | | | | | LegalBench-RAG | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | P↑ | R↑ | F1↑ | Ans.↑ | TR↑ | PB↑ | P↑ | R↑ | F1↑ | A-F1↑ | TR↑ | PB↑ | P↑ | R↑ | F1↑ | Ans.↑ | TR↑ | PB↑ |
| *Relevance-based retrieval and reranking* | | | | | | | | | | | | | | | | | | |
| BM25 | 28.79 | 14.27 | 15.14 | 59.38 | 2.40 | 62.50 | 43.24 | 78.29 | 51.44 | 43.24 | 19.87 | 68.66 | 11.69 | 13.22 | 11.15 | 19.08 | 1.50 | 20.06 |
| BGE-Reranker | 38.15 | 28.34 | 26.40 | 84.38 | 8.95 | 87.50 | 43.09 | 83.50 | 51.98 | 45.11 | 17.74 | 76.60 | 22.47 | 30.80 | 22.88 | 33.85 | 5.67 | 35.38 |
| MedCPT | 31.35 | 24.38 | 22.28 | 75.00 | 5.83 | 81.25 | – | – | – | – | – | – | – | – | – | – | – | – |
| Qwen3-Embedding | 38.11 | 27.48 | 25.52 | 86.46 | 0.00 | 82.29 | 44.77 | 83.09 | 53.49 | 44.90 | 0.00 | 78.09 | 25.13 | 31.56 | 25.43 | 34.77 | 0.00 | 36.92 |
| Qwen3-Reranker-4B | 31.31 | 32.63 | 26.23 | 82.29 | 12.49 | 88.54 | 43.33 | 85.15 | 52.37 | 44.77 | 19.23 | 78.37 | 24.15 | 37.86 | 26.25 | 41.23 | 13.44 | 43.83 |
| *Budget, redundancy, and context compression* | | | | | | | | | | | | | | | | | | |
| MMR | 31.17 | 17.43 | 17.68 | 79.17 | 0.77 | 77.08 | 45.48 | 83.46 | 54.32 | 45.07 | 3.21 | 75.35 | 17.42 | 20.13 | 16.46 | 27.69 | 1.10 | 27.08 |
| Adaptive-$k$ | **40.05** | 13.59 | 15.74 | 61.46 | 0.00 | 59.38 | 35.22 | 44.85 | 36.53 | 26.53 | 0.00 | 37.32 | 9.04 | 9.68 | 8.46 | 12.92 | 0.00 | 13.85 |
| RECOMP-extractive | 3.19 | 12.27 | 4.26 | 59.38 | 2.64 | 63.54 | 33.22 | 85.74 | 40.77 | 44.65 | 19.23 | 76.41 | 16.84 | 18.94 | 16.04 | 22.77 | 3.49 | 24.92 |
| Provence-pretrained | 12.98 | **46.64** | 17.35 | 86.46 | **33.96** | 87.50 | 11.17 | 88.84 | 18.17 | 41.67 | **23.73** | 85.16 | 1.57 | **42.89** | 2.93 | 42.15 | **23.61** | 47.37 |
| AdaGReS | 9.33 | 31.11 | 11.52 | 84.38 | 2.34 | 88.42 | 5.63 | **92.01** | 10.11 | 39.53 | 10.90 | 84.86 | 1.60 | 27.36 | 2.89 | 39.08 | 2.31 | 36.84 |
| *Listwise and set-based selection* | | | | | | | | | | | | | | | | | | |
| RankGPT-Qwen | 30.45 | 33.91 | 26.15 | 83.33 | 0.94 | 86.46 | 40.07 | 87.18 | 48.31 | 44.82 | 0.00 | 80.63 | 27.14 | 37.24 | 28.12 | 39.38 | 7.14 | 44.44 |
| SetR-prompt-Qwen | 39.36 | 26.36 | 26.19 | 82.29 | 0.00 | 85.42 | 45.16 | 83.79 | 54.19 | 45.14 | 0.00 | 79.86 | 22.66 | 34.82 | 24.18 | 39.08 | 0.00 | 41.98 |
| PACE-Frontloading | 29.79 | 10.51 | 12.02 | 75.00 | 2.04 | 73.03 | 44.22 | 68.31 | 49.58 | 39.39 | 3.21 | 54.61 | 7.84 | 8.55 | 6.92 | 20.92 | 0.15 | 15.08 |
| METEORA-Qwen | 8.85 | 9.89 | 7.12 | 63.54 | 8.28 | 56.04 | 15.98 | 46.13 | 21.69 | 28.70 | 0.51 | 39.44 | 0.55 | 5.06 | 0.94 | 14.15 | 2.86 | 14.46 |
| IRCoT-Qwen | 24.13 | 8.68 | 10.15 | 39.58 | 1.72 | 37.50 | 37.17 | 54.28 | 40.85 | 31.35 | 2.56 | 45.07 | 6.81 | 6.19 | 5.55 | 13.54 | 1.84 | 11.38 |
| **CLESO-IP** | 26.75 | 28.56 | 24.33 | 84.38 | 0.62 | 88.54 | 23.87 | 89.02 | 33.73 | 43.25 | 3.35 | 84.40 | 15.15 | 39.11 | 19.03 | **47.69** | 8.62 | 51.08 |
| **CLESO-CP** | 36.42 | 28.99 | **26.95** | **87.50** | 0.81 | **89.58** | **47.77** | 89.37 | **56.25** | **45.25** | 2.56 | **85.82** | **29.88** | 38.92 | **29.98** | 46.46 | 8.62 | **53.40** |

* *MedCPT is evaluated only on BioASQ. Values are question-macro means over valid outcomes.*

Table 3: The method F1 score, time, and token cost* across three benchmarks at $B_{\text{obs}} = 3{,}250$.**

| Method | BioASQ | | | QASPER | | | LegalBench-RAG | | |
|---|---|---|---|---|---|---|---|---|---|
| | F1↑ | Time s/q↓ | Token k/q↓ | F1↑ | Time s/q↓ | Token k/q↓ | F1↑ | Time s/q↓ | Token k/q↓ |
| BM25 | 15.14 | 0.39 | – | 51.44 | 0.0047 | – | 11.15 | 0.88 | – |
| BGE-Reranker-v2-m3 | 26.40 | 129.29 | 1629.4 | 51.98 | 0.83 | 8.8 | 22.88 | 653.58 | 4120.0 |
| Qwen3-Reranker-4B | 26.23 | 265.57 | 3505.3 | 52.37 | 1.90 | 20.5 | 26.25 | 1168.93 | 6459.7 |
| Provence-pretrained | 17.35 | 146.53 | 891.7 | 18.17 | 0.81 | 5.7 | 2.93 | 300.87 | 2514.3 |
| RankGPT-Qwen | 26.15 | 6.54 | 11.6 | 48.31 | 58.76 | 10.5 | 28.12 | 79.22 | 21.0 |
| SetR-prompt-Qwen | 26.19 | 6.09 | 10.7 | 54.19 | 14.63 | 9.5 | 24.18 | 26.35 | 7.6 |
| METEORA-Qwen | 7.12 | 61.54 | 985.0 | 21.69 | 27.81 | 6.6 | 0.94 | 106.67 | 2337.8 |
| IRCoT-Qwen | 10.15 | 2.20 | 30.7 | 40.85 | 10.63 | 34.2 | 5.55 | 23.76 | 46.2 |
| **CLESO-IP** | 24.33 | 72.86 | 83.3 | 33.73 | 93.15 | 76.9 | 19.03 | 52.94 | 64.2 |
| **CLESO-CP** | **26.95** | 29.73 | 71.4 | **56.25** | 25.54 | 41.9 | **29.98** | 16.59 | 53.6 |

* Green (red) time and input-token values are lower(higher) than CLESO-CP on the same benchmark.
** Time is serial-equivalent; Input is cumulative model-input tokens over recorded calls. CLESO uses Qwen Reranker from the default 4B.

### 6.2 Main Experiment (RQ1)

We compare methods on the full Test splits under a common 3,250-source-token cap on final selected evidence. For CLESO, the same cap also limits the source text admitted to its observation pool, while Probe has an independent budget. Ranking-based methods use a shared LLM selector. Table 2 reports final-evidence and answer metrics. CLESO-CP obtains the highest reported evidence F1 in all three benchmarks: 26.95%, 56.25%, and 29.98%. These values exceed the strongest reported baseline means by 0.55, 1.93, and 1.86 percentage points, respectively. CP also has the highest Answer score on BioASQ and QASPER; on LegalBench-RAG, CLESO-IP has the highest Answer score (47.69%, versus 46.46% for CP). CP achieves the highest reference-conditioned PB in all three benchmarks (89.58%, 85.82%, and 53.40%) and CP's F1 advantage over IP is largest on QASPER, although PB is a model-assisted support measure rather than a test of answer stability. Tail Recall follows a different pattern: Provence has the highest value in each benchmark. Thus, the evidence-F1 advantage does not imply uniformly better recovery of tail targets.

### 6.3 Budget and Cost Scalability (RQ2)

Table 3 and Fig. 1(B) report the full-Test evidence-quality and method-native cost profile of CLESO with a 4B package reranker. At $B_{\text{obs}} = 3{,}250$, CLESO-CP achieves stronger evidence F1 than CLESO-IP with lower recorded runtime and cumulative model input on all three benchmarks.

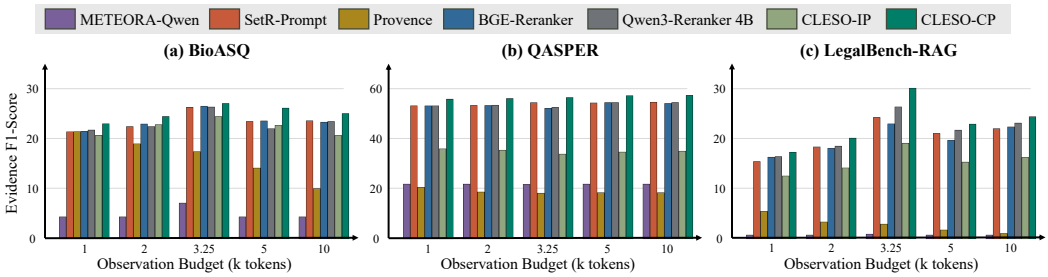


Figure 3: Full-Test Evidence F1 at five observation budgets for CLESO-IP/CP and five baselines.

On LegalBench-RAG, CP uses 16.59 seconds and 53.6k model-input tokens per question, versus 1,168.93 seconds and 6.46M tokens for Qwen3-Reranker. The recorded cost advantage over full-queue reranking is most pronounced on the long-sequence BioASQ and LegalBench-RAG benchmarks. The observation cap bounds the source text admitted to $O$, not cumulative model input.

Fig. 3 varies the observation budget from 1k to 10k source tokens on the test cases while holding CLESO's 4B package scorer fixed. CLESO-CP has the highest Evidence F1 among the seven plotted methods in all 15 benchmark-budget settings. At 3.25k tokens, it reaches 26.95%, 56.25%, and 29.98% on BioASQ, QASPER, and LegalBench-RAG, respectively, and it exceeds CLESO-IP at every tested budget. The advantage persists across all tested budgets on the two long-queue benchmarks. More observation does not always improve final selection: BioASQ and LegalBench-RAG peak at 3.25k tokens among the tested settings, whereas QASPER rises to 57.12% at 10k tokens.

### 6.4 Structural Transfer and Ablations (RQ3)

Before Probe observation and package selection, the fixed student used in the full-Test experiments attains 3,250-token Dev global-prefix target recall of 50.70%, 85.12%, and 35.28% on BioASQ, QASPER, and LegalBench-RAG. The original-ranking values are 49.37%, 85.04%, and 34.74%, whereas the label-visible criterion references reach 73.88%, 88.63%, and 75.02%. Thus, final CP gains do not require exact reproduction of reference-prefix access.

Table 4 separates target exposure from final evidence selection. Under FLAT packing, replacing RAW access with interval access changes final precision by less than one point in each benchmark; intervalization alone does not explain the final precision gain. The unpacked Probe pool reaches target recall of 34.78%, 94.51%, and 41.49% on BioASQ, QASPER, and LegalBench-RAG. With CP fixed, Probe allocation raises final recall over equal interval quotas by 21.66, 1.74, and 35.36 points. With the observation pool fixed, CP raises precision over IP by 9.67, 23.90, and 14.73 points, while recall changes by at most 0.43 points. This comparison supports CP's depth-wise packaging policy but does not isolate cross-interval composition from package granularity.

Table 4: Ablation study in final-evidence selection (%)*.

| Variant | BioASQ | | | QASPER | | | LB-RAG | | |
|---|---|---|---|---|---|---|---|---|---|
| | P↑ | R↑ | Ans.↑ | P↑ | R↑ | Ans.↑ | P↑ | R↑ | Ans.↑ |
| RAW + FLAT | 23.61 | 31.41 | 85.26 | 24.33 | 88.52 | 42.48 | 14.33 | 35.32 | 40.92 |
| Int. + FLAT | 22.95 | 31.63 | 85.42 | 23.78 | 88.64 | 43.08 | 14.49 | 39.68 | 44.92 |
| Int. + EB + CP | 24.01 | 7.33 | 55.21 | 45.67 | 87.63 | 45.18 | 4.04 | 3.56 | 10.46 |
| Int. + Probe (unpack) | 8.54 | **34.78** | 85.42 | 4.60 | **94.51** | 41.47 | 1.65 | **41.49** | 44.92 |
| Int. + Probe + FLAT | 26.43 | 28.56 | 84.38 | 24.07 | 88.71 | 43.43 | 14.96 | 38.74 | 46.46 |
| Int. + Probe + IP | 26.75 | 28.56 | 84.38 | 23.87 | 89.02 | 43.25 | 15.15 | 39.11 | **47.69** |
| Int. + Probe + CP | **36.42** | 28.99 | **87.50** | **47.77** | 89.37 | **45.25** | **29.88** | 38.92 | 46.46 |

* EB (equal budget) assigns uniform observation quotas to intervals; FLAT greedily packs observed occurrences in presentation order without interval traversal.

## 7 Conclusion

In this work, we introduced CLESO for observation-budgeted probative evidence selection in domain-specific QA. Ontology-guided criterion search organizes local support priorities and separable intervals, which a lightweight refiner learns from inference-visible inputs. Recovered intervals guide Probe allocation and local-head observation; proposal-conditioned packing then enables joint evidence selection. At the matched 3,250-token setting, CLESO-CP attains the highest Evidence F1 on all three benchmarks and the highest Answer score on BioASQ and QASPER; CLESO-IP has the highest LegalBench-RAG Answer score. Ablations link Probe allocation to target exposure and CP's depth-wise packing to final-set precision. Our analysis specifies conditions for interval recovery and budgeted witness exposure without equating exposure with final probative selection.

## AI USE STATEMENT

This work used generative AI tools to assist with searching for, collecting, and summarizing relevant literature; drafting and editing supplementary and other non-main-text material; polishing the completed main-text draft before final human review; and supporting the experimental workflow, including feedback on experimental design, assistance with implementation and execution, and interpretation of results. The authors reviewed AI-assisted text, code, and experimental outputs, checked cited sources and reported results against the underlying materials and experiment records, and made all final methodological, analytical, and editorial decisions.

## REFERENCES

Akari Asai, Timo Schick, Patrick Lewis, Xilun Chen, Gautier Izacard, Sebastian Riedel, Hannaneh Hajishirzi, and Wen-tau Yih. Task-aware retrieval with instructions. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, volume ACL 2023 of *Findings of ACL*, pp. 3650–3675. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.FINDINGS-ACL.225. URL https://doi.org/10.18653/v1/2023.findings-acl.225.

Weibin Cai and Reza Zafarani. Less can be more: Relieving RAG bottlenecks via evidence frontloading and pressure-adaptive budgeting, 2026. URL https://doi.org/10.48550/arXiv.2608.25115.

Nadezhda Chirkova, Thibault Formal, Vassilina Nikoulina, and Stéphane Clinchant. Provence: efficient and robust context pruning for retrieval-augmented generation. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=TDy5Ih78b4.

Computer Science Ontology. Computer Science Ontology v3.5, July 2025. URL https://cso.kmi.open.ac.uk/downloads. Released 29 July 2025. Accessed 25 September 2026.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. A dataset of information-seeking questions and answers anchored in research papers. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tür, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pp. 4599–4610. Association for Computational Linguistics, 2021. doi: 10.18653/V1/2021.NAACL-MAIN.365. URL https://doi.org/10.18653/v1/2021.naacl-main.365.

Data Privacy Vocabularies and Controls Community Group. Data Privacy Vocabulary (DPV), Version 2.3, February 2026. URL https://w3c-cg.github.io/dpv/2.3/dpv/. Final Community Group Report, 25 February 2026. Accessed 25 September 2026.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. From local to global: A graph RAG approach to query-focused summarization. *CoRR*, abs/2404.16130, 2024. doi: 10.48550/ARXIV.2404.16130. URL https://doi.org/10.48550/arXiv.2404.16130.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. A survey on RAG meeting llms: Towards retrieval-augmented large language models. In Ricardo Baeza-Yates and Francesco Bonchi (eds.), *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2024, Barcelona, Spain, August 25-29, 2024*, pp. 6491–6501. ACM, 2024. doi: 10.1145/3637528.3671470. URL https://doi.org/10.1145/3637528.3671470.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate text with citations. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pp. 6465–6488. Association for Computational Linguistics, 2023. doi:

10.18653/V1/2023.EMNLP-MAIN.398. URL https://doi.org/10.18653/v1/2023.emnlp-main.398.

Jade Goldstein and Jaime G. Carbonell. Summarization: (1) using MMR for diversity- based reranking and (2) evaluating summaries. In *TIPSTER TEXT PROGRAM PHASE III: Proceedings of a Workshop held at Baltimore, MD, USA, October 13-15, 1998*, pp. 181–195. Morgan Kaufmann, 1998. doi: 10.3115/1119089.1119120. URL https://aclanthology.org/X98-1025/.

Kailin Jiang, Lei Liu, Jian Xi, Hui Xu, Junlin Liu, Baochen Fu, Bin Li, Vichwang, Yu Lu, and Haibo Shi. Beyond relevance-centric retrieval: Rubric-oriented document set selection and ranking, 2026. URL https://doi.org/10.48550/arXiv.2607.19747.

Qiao Jin, Won Kim, Qingyu Chen, Donald C. Comeau, Lana Yeganova, W. John Wilbur, and Zhiyong Lu. Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. *Bioinform.*, 39(10), 2023. doi: 10.1093/BIOINFORMATICS/BTAD651. URL https://doi.org/10.1093/bioinformatics/btad651.

Ansh Kamthan. AB-RAG: adaptive budgeted retrieval-augmented generation for reliable question answering, 2026. URL https://doi.org/10.48550/arXiv.2606.29090.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pp. 6769–6781. Association for Computational Linguistics, 2020. doi: 10.18653/V1/2020.EMNLP-MAIN.550. URL https://doi.org/10.18653/v1/2020.emnlp-main.550.

Omar Khattab and Matei Zaharia. Colbert: Efficient and effective passage search via contextualized late interaction over BERT. In Jimmy X. Huang, Yi Chang, Xueqi Cheng, Jaap Kamps, Vanessa Murdock, Ji-Rong Wen, and Yiqun Liu (eds.), *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pp. 39–48. ACM, 2020. doi: 10.1145/3397271.3401075. URL https://doi.org/10.1145/3397271.3401075.

Daegun Lee, Jiwoo Choi, Gyeong Hoon Yi, Seok Su Sohn, Byungju Lee, and Donghun Kim. Structured domain knowledge enables trustworthy materials science question-answering with large language models. *Digital Discovery*, 5:2243–2253, 2026. doi: 10.1039/D6DD00028B. URL https://doi.org/10.1039/D6DD00028B.

Dahyun Lee, Yongrae Jo, Haeju Park, and Moontae Lee. Shifting from ranking to set selection for retrieval augmented generation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pp. 17606–17619. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.ACL-LONG.861. URL https://doi.org/10.18653/v1/2025.acl-long.861.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

Minghan Li, Junjie Zou, Xinxuan Lv, Chao Zhang, and Guodong Zhou. S2G-RAG: structured sufficiency and gap judging for iterative retrieval-augmented QA. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2026, San Diego,*

*California, United States, July 2-7, 2026*, pp. 25846–25862. Association for Computational Linguistics, 2026. doi: 10.18653/V1/2026.ACL-LONG.1185. URL https://doi.org/10.18653/v1/2026.acl-long.1185.

Zhenpeng Li. Budgeted evidence set construction for cross-library compliance RAG: Beyond relevance ranking. SSRN, 2026. URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=6987087.

Zhuoqun Li, Xuanang Chen, Haiyang Yu, Hongyu Lin, Yaojie Lu, Qiaoyu Tang, Fei Huang, Xianpei Han, Le Sun, and Yongbin Li. Structrag: Boosting knowledge intensive reasoning of llms via inference-time hybrid information structurization. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=GhexuBLxbO.

Xiangyang Mou, Mo Yu, Shiyu Chang, Yufei Feng, Li Zhang, and Hui Su. Complementary evidence identification in open-domain question answering. In Paola Merlo, Jörg Tiedemann, and Reut Tsarfaty (eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pp. 2720–2726. Association for Computational Linguistics, 2021. doi: 10.18653/V1/2021.EACL-MAIN.234. URL https://doi.org/10.18653/v1/2021.eacl-main.234.

noslegal. noslegal taxonomy v4, May 2026. URL https://github.com/noslegal/taxonomy. Released under the Apache 2.0 licence. Accessed 25 September 2026.

Chao Peng, Bin Wang, Zhilei Long, and Jinfang Sheng. Adagres:adaptive greedy context selection via redundancy-aware scoring for token-budgeted RAG, 2025. URL https://doi.org/10.48550/arXiv.2512.25052.

Nicholas Pipitone and Ghita Houir Alami. Legalbench-rag: A benchmark for retrieval-augmented generation in the legal domain. *CoRR*, abs/2408.10343, 2024. doi: 10.48550/ARXIV.2408.10343. URL https://doi.org/10.48550/arXiv.2408.10343.

Sashank J. Reddi, Rama Kumar Pasumarthi, Aditya Krishna Menon, Ankit Singh Rawat, Felix X. Yu, Seungyeon Kim, Andreas Veit, and Sanjiv Kumar. Rankdistil: Knowledge distillation for ranking. In Arindam Banerjee and Kenji Fukumizu (eds.), *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021, April 13-15, 2021, Virtual Event*, volume 130 of *Proceedings of Machine Learning Research*, pp. 2368–2376. PMLR, 2021. URL http://proceedings.mlr.press/v130/reddi21a.html.

Stephen E. Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389, 2009. doi: 10.1561/1500000019. URL https://doi.org/10.1561/1500000019.

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D. Manning. RAPTOR: recursive abstractive processing for tree-organized retrieval. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=GN921JHCRw.

Yash Saxena, Ankur Padia, Mandar S. Chaudhary, Kalpa Gunaratna, Srinivasan Parthasarathy, and Manas Gaur. Ranking free RAG: replacing re-ranking with selection in RAG for sensitive domains, 2025. URL https://doi.org/10.48550/arXiv.2505.16014.

David Shotton and Silvio Peroni. The Discourse Elements Ontology (DEO), Version 1.2.0, August 2026. URL https://sparontologies.github.io/deo/current/deo.html. SPAR Ontologies. Version dated 14 August 2026. Accessed 25 September 2026.

Rahul Singh and Madhav Vadlamani. Qubo-optimized evidence selection for retrieval-augmented question answering with unconventional solvers, 2026. URL https://doi.org/10.48550/arXiv.2607.12334.

Yongho Song, Dahyun Lee, Myungha Jang, Seung-won Hwang, Kyungjae Lee, Dongha Lee, and Jinyoung Yeo. Evidentiality-aware retrieval for overcoming abstractiveness in open-domain question answering. In Yvette Graham and Matthew Purver (eds.), *Findings of the Association for*

*Computational Linguistics: EACL 2024, St. Julian's, Malta, March 17-22, 2024*, volume EACL 2024 of *Findings of ACL*, pp. 1930–1943. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.FINDINGS-EACL.130. URL https://doi.org/10.18653/v1/2024.findings-eacl.130.

Jiashuo Sun, Pengcheng Jiang, Saizhuo Wang, Jiajun Fan, Heng Wang, Siru Ouyang, Ming Zhong, Yizhu Jiao, Chengsong Huang, Xueqiang Xu, Pengrui Han, Peiran Li, Jiaxin Huang, Ge Liu, Heng Ji, and Jiawei Han. Rethinking the reranker: Boundary-aware evidence selection for robust retrieval-augmented generation, 2026. URL https://doi.org/10.48550/arXiv.2602.03689.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. Is chatgpt good at search? investigating large language models as re-ranking agents. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pp. 14918–14937. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.EMNLP-MAIN.923. URL https://doi.org/10.18653/v1/2023.emnlp-main.923.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pp. 10014–10037. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.ACL-LONG.557. URL https://doi.org/10.18653/v1/2023.acl-long.557.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R. Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, Yannis Almirantis, John Pavlopoulos, Nicolas Baskiotis, Patrick Gallinari, Thierry Artières, Axel-Cyrille Ngonga Ngomo, Norman Heino, Éric Gaussier, Liliana Barrio-Alvers, Michael Schroeder, Ion Androutsopoulos, and Georgios Paliouras. An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. *BMC Bioinform.*, 16:138:1–138:28, 2015. doi: 10.1186/S12859-015-0564-6. URL https://doi.org/10.1186/s12859-015-0564-6.

Victor P. Unda. Controllable evidence selection in retrieval-augmented question answering via deterministic utility gating, 2026. URL https://doi.org/10.48550/arXiv.2603.18011.

U.S. National Library of Medicine. 2026AA UMLS Release Notes and Bugs, May 2026. URL https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/release/notes.html. UMLS Knowledge Sources, 2026AA release. Last reviewed May 4, 2026. Accessed September 25, 2026.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 5998–6008, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Orion Weller, Benjamin Chang, Sean MacAvaney, Kyle Lo, Arman Cohan, Benjamin Van Durme, Dawn J. Lawrie, and Luca Soldaini. Followir: Evaluating and teaching information retrieval models to follow instructions. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pp. 11926–11942. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.NAACL-LONG.597. URL https://doi.org/10.18653/v1/2025.naacl-long.597.

Mingyan Wu, Han Yang, Omer Ben-Porat, and Yftah Ziser. When knowledge is not free: Cost-aware evidence selection in retrieval-augmented generation, 2026a. URL https://doi.org/10.48550/arXiv.2606.02245.

Yuhang Wu, Xiangqing Shen, Fanfan Wang, Cangqi Zhou, Zhen Wu, Xinyu Dai, and Rui Xia. Optimizing RAG rerankers with LLM feedback via reinforcement learning. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2026, San Diego, California, United States, July 2-7, 2026*, pp. 30474–30490. Association for Computational Linguistics, 2026b. doi: 10.18653/V1/2026.ACL-LONG.1406. URL https://doi.org/10.18653/v1/2026.acl-long.1406.

Fangyuan Xu, Weijia Shi, and Eunsol Choi. RECOMP: improving retrieval-augmented lms with context compression and selective augmentation. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=mlJLVigNHp.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 embedding: Advancing text embedding and reranking through foundation models, 2025. URL https://doi.org/10.48550/arXiv.2506.05176.

# APPENDIX

## A CASE STUDIES OF EVIDENCE ROLES AND BUDGETED OBSERVATION

We present one case per benchmark to illustrate two challenges: identifying answer-bearing roles among related candidates and observing dispersed support within a fixed source-text budget. The cases come from the project's LegalBench-RAG sentence pool, QASPER's training partition, and the exposed BioASQ fit260 training subset. They were selected for source-visible role contrasts rather than CLESO outcomes. Green marks statements contributing support; red marks related text that fails to establish the required relation or source binding, without implying factual falsity.

All ranks refer to candidate occurrences in the frozen relevance ordering, not document ranks. LegalBench-RAG uses the project's sentence pools from 50 retrieved documents; QASPER uses sentences from the question's source paper. Their rankings were replayed from frozen RAW vectors, while BioASQ uses the existing RAW rankings. Displayed excerpts are aligned in the source audit, with whitespace normalized. Token costs use the frozen tokenizer and charge complete candidates, excluding prompt framing; a prefix stops before the first candidate that would exceed the budget. Support costs below describe the designated statements, not globally minimal sufficient sets. These are qualitative source analyses, not fixed-answerer Probative evaluations.

### A.1 CASE STUDY OF LEGALBENCH-RAG

**Case 1: A destruction obligation with a displaced trigger (QID:***8612f2)**

**Question.** Consider eHandshake's Non-Disclosure Agreement; Does the document specify whether the Receiving Party is required to destroy or return Confidential Information upon the termination of the Agreement?

**Related but insufficient.**

***Rank 1**, UW Oshkosh NDA, a different agreement:*

"The Receiving Party will return or destroy Confidential Information provided by the Disclosing Party upon termination of the Agreement."

**Supporting evidence.**

***Rank 53**, the same agreement, destruction duty:*

"Upon receipt of such communication, the Receiving Party shall destroy promptly all documents furnished and will not retain any copies, extracts or other reproductions in whole or in part of such Information."

***Rank 218**, the same agreement, retention qualification:*

"Notwithstanding the foregoing, the Receiving Party shall be entitled to retain that portion of the Information for legal, regulatory or internal compliance purposes, which will continue to be treated as confidential on the terms hereof."

***Rank 570**, eHandshake, decision and notice:*

"If any of the Parties determine that they do not wish to proceed or continue with the Business Under Discussion, it will promptly advise the other Party of that decision in writing or by email."

**Role contrast and observation consequence.** The rank-1 sentence matches the requested obligation but belongs to another NDA. In eHandshake, the destruction duty depends on "such communication," whose antecedent specifies a written or emailed decision not to proceed or continue with the business under discussion. The supported reading is therefore a notice-triggered duty with a retention qualification, rather than an unconditional obligation upon every form of termination. The trigger and duty cost 73 tokens, or 115 tokens including the retention clause, but their RAW acquisition prefix reaches 36,672 tokens. Neither the duty nor its trigger is present within the 3,250-token prefix. This case combines source-identity confusion with the separation of a statement from the context that determines when it applies. This is a source-grounded contract reading; the audited materials do not provide an official answer reference for this case.

### A.2 CASE STUDY OF QASPER

**Case 2: Evaluation success versus the measures being evaluated (QID: ***c6c90b)**

**Question.** What automatic and human evaluation metrics are used to compare SPNet to its counterparts?

**Related but insufficient.**

***Rank 1**, performance conclusion:*

"We observe that SPNet outperforms other methods in all the automatic evaluation metrics with a big margin, as it incorporates all three semantic scaffolds."

**Supporting evidence.**

***Rank 4**, automatic measures:*

"We observe that SPNet reaches the highest score in both ROUGE and CIC."

***Rank 36**, the systems being compared by human participants:*

"For each sample, we show the conversation, reference summary, as well as summaries generated by Pointer-Generator and SPNet to three different participants."

***Rank 50**, human measures and evaluation procedure:*

"The participants are asked to score each summary on three indicators: relevance, conciseness, and readability on a 1 to 5 scale, and rank the summary pair (ties allowed)."

**Role contrast and observation consequence.** The rank-1 concerns SPNet and its evaluation, but reports a performance comparison without identifying the measures. Answering the question requires the automatic metric names and the human criteria, with the participant procedure bound to SPNet and its comparator. These complementary roles occur within one paper, rather than across independent studies. The three supporting candidates cost 82 tokens and are all available after 1,063 RAW-prefix tokens. This case isolates evidence-role disentanglement, not a coverage failure at 3,250 tokens: relevant evaluation language must be distinguished from the specific information requested.

### A.3 Case Study of BioASQ

**Case 3: Method capability versus correctly attributed advantages (QID:***00007d)**

**Question.** What are the main advantages of mirnovo compared to existing tools?

**Related but insufficient.**

***Rank 5**, PMID 36437740, a speed comparison for miRador:*

"We compared miRador to other commonly used miRNA prediction tools and found that miRador is at least as precise as other prediction tools while being substantially faster than other tools."

***Rank 11**, PMID 33381847, a usability claim for MirCure:*

"MirCure is an easy-to-use tool with a graphical interface that allows both scoring of miRNA reliability and browsing of supporting evidence by manual curators."

**Supporting evidence.**

***Rank 1**, PMID 29036314, mirnovo identity and capability:*

"To this end, we have developed mirnovo, a machine learning-based algorithm, which can identify known and novel miRNAs in animals and plants directly from small RNA-Seq data, with or without a reference genome."

***Rank 5,058**, the same abstract, comparative result:*

"This method performs comparably to existing tools; however, it is simpler to use with reduced run time."

**Role contrast and observation consequence.** The rank-1 mirnovo description establishes its ability to operate without a reference genome, but not comparative speed or usability, which early passages attribute to miRador and MirCure, respectively. The mirnovo abstract reports comparable performance, simpler use, and reduced run time at rank 5,058. Although adjacent in that abstract, this sentence and its rank-1 antecedent are widely separated in the relevance queue. The displayed pair contains 66 tokens, yet reaching both through the RAW prefix costs 143,906 tokens, leaving the comparative benefits outside the reviewed 3,250- and 4,000-token prefixes despite the useful partial information they contain. The first statement resolves the referent of "This method", so their pairing preserves source binding rather than adding a second independent answer requirement.

# B FORMALIZATION OF BUDGETED PROBATIVE EVIDENCE SELECTION

This appendix specifies the source-grounding and budget constraints of the task. It also distinguishes between exposing a probative evidence set and successfully selecting one. Method details and conditional access guarantees appear in Appendices C and D.

## B.1 SOURCE-GROUNDED CASES AND PROBATIVE EVIDENCE

A case is $x = (q, P)$, where $P = (p_1, \ldots, p_n)$ is the relevance-ranked candidate sequence with original scores $\mathbf{s}$ and order $\pi_P$. Its supervision is $y = (E^*, a^*)$. Let $V_x = \{1, \ldots, n\}$ index candidate occurrences. Each occurrence retains its identity and source coordinates.

**Definition 3** (Source-grounded evidence and observation)**.** *The candidate-grounded span universe and finite evidence-set space are*

$$\mathcal{S}(P) = \{e : \exists p \in P,\ e \subseteq_{\text{char}} p\}, \quad \mathcal{E}_x = \{E \subseteq \mathcal{S}(P) : |E| < \infty\}. \tag{11}$$

*An observation $O$ is an ordered collection of complete candidate occurrences with distinct occurrence identities. Let $\chi(E)$ be the union of source-character positions of an evidence set $E$, distinguished by source identity, and let $\chi(O)$ be the corresponding union for the occurrences in $O$. We write $E \preceq O$ when $\chi(E) \subseteq \chi(O)$.*

**Definition 4** (Probative evidence set)**.** *Fix an answerer $L$, its evidence serialization, an answer-equivalence relation $\equiv$, and stochastic execution conditions $(\omega_t)_{t\geq 1}$. An evidence set $E \in \mathcal{E}_x$ is probative for $(x, y)$ if, for every $\epsilon > 0$, there exists $N \in \mathbb{N}$ such that, for every $T > N$,*

$$\left| \tfrac{1}{T} \textstyle\sum_{t=1}^{T} \mathbf{1}\{L(q, E; \omega_t) \equiv a^*\} - 1 \right| < \epsilon. \tag{12}$$

*Equivalently,*

$$\mathrm{Rel}_{x,y}(E) = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbf{1}\{L(q, E; \omega_t) \equiv a^*\}, \quad \mathcal{U}_{x,y} = \{E \in \mathcal{E}_x : \mathrm{Rel}_{x,y}(E) = 1\}. \tag{13}$$

*The definition applies when the displayed limit exists. Eq. 12 requires it to exist and equal one.*

As assumed in the Preliminary, the annotated set $E^*$ is one witness in $\mathcal{U}_{x,y}$. Other source-grounded sets may also be probative, so successful selection need not reproduce $E^*$ exactly. Reliability is defined for a complete evidence set; adding source content is not assumed to improve it monotonically.

## B.2 BUDGET-FEASIBLE OBSERVATION AND SELECTION

**Definition 5** (Observation and selection policies)**.** *An observation policy $g$ constructs an ordered collection $O = g(x)$ of complete candidate occurrences. A selection policy $f$ maps the question and observation to an evidence set:*

$$O = g(x), \qquad \hat{E} = f(q, O) \in \mathcal{E}_x, \qquad \hat{E} \preceq O. \tag{14}$$

*Both policies use inference-available information and do not receive $y = (E^*, a^*)$.*

Reference evidence, reference answers, fitted criteria, and criterion-derived structures may supervise training but are not additional deployment inputs. Full-queue feature construction and model execution have costs separate from the source-text observation budget.

**Definition 6** (Budget-feasible policy pair)**.** *Let $A_O \subseteq V_x$ contain the occurrence identities admitted to $O$. For the fixed tokenizer, let $c_i > 0$ be the charged source-text token cost of $p_i$. The observation cost is additive over admitted occurrences: $\ell(O) = \sum_{i\in A_O} c_i$.*

*For budget $B_{\text{obs}}$, feasible policy pairs satisfy*

$$\Pi(B_{\text{obs}}) = \left\{ (g, f) : \begin{array}{l} \ell(g(x)) \leq B_{\text{obs}}, \\ f(q, g(x)) \in \mathcal{E}_x, \quad f(q, g(x)) \preceq g(x), \quad \forall x \in \mathcal{X} \end{array} \right\}. \tag{15}$$

*For randomized policies, these constraints hold for every realized execution.*

Distinct occurrences are charged separately even if their source characters overlap; $\chi(O)$ counts overlapping source positions only once. Repeated model processing incurs additional computation without enlarging the observation pool. There is no separate cardinality or output budget on $\hat{E}$.

**Definition 7** (Observation-budgeted probative evidence selection)**.** *Given $\mathcal{D}$, $L$, and $B_{\mathrm{obs}}$, the task is to learn a feasible policy pair maximizing the probability of selecting a probative set. An optimal pair, when attained, satisfies*

$$(g^*_{B_{\mathrm{obs}}}, f^*_{B_{\mathrm{obs}}}) \in \operatorname*{arg\,max}_{(g,f)\in\Pi(B_{\mathrm{obs}})} \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathbf{1}\{f(q, g(x)) \in \mathcal{U}_{x,y}\}\right]. \tag{16}$$

*The expectation includes policy randomness when present. Success is membership in $\mathcal{U}_{x,y}$, rather than exact recovery of $E^*$.*

CLESO uses an additional, independent probing budget $B_{\mathrm{probe}}$. Let $H_g(x)$ denote the ordered collection read for probing before allocation. The two content constraints are

$$\ell(H_g(x)) \le B_{\mathrm{probe}}, \quad \ell(g(x)) \le B_{\mathrm{obs}}. \tag{17}$$

The probe informs allocation but does not automatically add its text to $O$. These budgets restrict source content at different stages; neither bounds cumulative model input, repeated processing, nor runtime.

An answer proposal $\tilde{a}$ generated from $(q, O)$ is an internal input to $f$, not reference supervision or an additional evidence source. The final answerer receives $(q, \hat{E})$ without the proposal or unselected observations.

### B.3 Observable Support and Selection Success

**Definition 8** (Observable probative evidence family)**.** *For a realized observation $O$, define*

$$\mathcal{U}_{x,y}(O) = \{E \in \mathcal{U}_{x,y} : E \preceq O\}. \tag{18}$$

*The observation admits probative selection when $\mathcal{U}_{x,y}(O) \neq \varnothing$.*

Exposing a witness $E^\dagger \in \mathcal{U}_{x,y}$ makes this family nonempty. The observation itself may contain distracting material and need not be probative.

**Proposition 9** (Observation and selection failure)**.** *For any feasible pair, let $O = g(x)$ and $\hat{E} = f(q, O)$. Then*

$$\Pr[\hat{E} \notin \mathcal{U}_{x,y}] = \Pr[\mathcal{U}_{x,y}(O) = \varnothing] + \Pr[\mathcal{U}_{x,y}(O) \neq \varnothing,\ \hat{E} \notin \mathcal{U}_{x,y}]. \tag{19}$$

*Consequently,* $\Pr[\hat{E} \in \mathcal{U}_{x,y}] \le \Pr[\mathcal{U}_{x,y}(O) \neq \varnothing]$. *Probabilities include cases and policy randomness.*

*Proof.* Feasibility gives $\hat{E} \preceq O$. If $\mathcal{U}_{x,y}(O)$ is empty, $\hat{E}$ cannot be probative. Otherwise, failure is attributable to selection despite observable support. These disjoint events partition the failure event, giving the equality and inequality. □

The decomposition separates access from selection. Criterion-derived supervision, score-based interval recovery, and Probe allocation determine which source content CLESO exposes. Proposal-conditioned packing restricts the package combinations available to the frozen LLM selector; successful selection further depends on whether a probative combination is expressible and chosen. Section 5 analyzes access and budget costs, with proofs in Appendix D.

Source-overlap scores, finite answer samples, and reference-conditioned support judgments provide operational evaluations of these stages. They do not replace membership in the ideal family $\mathcal{U}_{x,y}$; their protocols are specified in Appendix E.

# C METHOD DETAILS OF CRITERION-GUIDED SEQUENCE ACCESS

This appendix describes how criterion-supported regions become score supervision and how inference-time scores control budgeted observation and package selection. It uses the ordered occurrence observation and additive source-token accounting of Appendix B.

## C.1 CRITERION CONSTRUCTION AND REFERENCE SEQUENCE

**Inputs and target identities.** Each occurrence $p_i$ retains its candidate identity, source identifier, character offsets, and complete text. With fixed ontology resources $\Omega$, the incidence matrix contains predicates extracted from the question and candidates:

$$M_x(i, \alpha) = \mathbf{1}\{\alpha \in \Gamma_\Omega(q, p_i)\}. \tag{20}$$

The extraction procedure and ontology resources are fixed across splits. Missing extraction is not treated as a verified negative predicate. BioASQ uses grounded biomedical concepts and relations; QASPER and LegalBench-RAG use domain schemas and source-bound attributes. During training, positive source-character overlap with $E^*$ induces candidate-level target identities. Neither these identities nor fitted criteria enter the inference-time feature vector.

**Forward search and partition.** The training-time search evaluates the AND–OR–VETO criterion $R(\eta)$ defined in the Approach. Its bounded executable family permits at most five clauses and three positive branches; the cross-domain forward search uses up to 256 required atoms and beam width 1,024. Starting at the first unprocessed target in the original relevance order, it extends through consecutive target positions while a criterion exactly matches the targets in the proposed region. At the first failed extension, it retains the last feasible region; a target without a feasible starting rule is recorded as uncovered. Accepted matches retain candidate identities and source bindings as local hyperedges. These records establish supervision provenance; the refiner does not use a separate hyperedge-completion loss.

Original-score gaps refine cuts within ranges that preserve the fitted target group and avoid rule-selected non-targets. Remaining prefix, inter-region, and suffix candidates complete the ordered partition $\mathcal{I}_x^{\mathrm{cs}}$ of contiguous original-position intervals. Exact criterion fit applies to each searched region, not necessarily to every candidate subsequently assigned to its completed interval. If no region is feasible, the complete sequence forms one fallback interval. This disjoint partition is the laminar case used by CLESO; semantic support may still cross intervals.

**Local priorities.** Within a completed interval, criterion-covered targets receive head priority (role 3). Non-targets satisfying a positive rule and a target-safe VETO receive tail priority (role 0). Other candidates, including uncovered targets, remain in the middle (role 1). BioASQ uses its stored positive/veto rule and cached forward regions; QASPER and LegalBench-RAG construct an auxiliary positive/veto rule for tail supervision, distinct from the exact rule fitted to a target group. Candidates are stably sorted by priority within each interval, preserving original order within a role. Write $\pi^{\mathrm{cs}}$ for the resulting permutation.

**Reference-score geometry.** Let $\mu_x = \mathrm{median}(\mathbf{s})$ and $a_x = \max\{\mathrm{IQR}(\mathbf{s}), 10^{-4}\}$. Along $\pi^{\mathrm{cs}}$, adjacent normalized score drops take one of three values: $d_{\mathrm{out}}$ at interval boundaries, $d_{\mathrm{role}}$ at role transitions, and $d_{\mathrm{in}}$ elsewhere. To retain these distinctions after FP32 storage, set

$$d_{\mathrm{in}} = \max\{2{\times}10^{-5}, 2\zeta_x\},\ d_{\mathrm{role}} = \max\{5{\times}10^{-4}, 2\zeta_x\},\ d_{\mathrm{out}} = \max\{2{\times}10^{-3}, d_{\mathrm{role}} + 2\zeta_x\},$$

where $\zeta_x = \max\{8|\,\mathrm{spacing}_{32}(\max_i |s_i|)|/a_x, 10^{-7}\}$. For $v_t = -\sum_{u<t} d_u$, the reference score is the descending path shifted to match the mean normalized original score:

$$s^{\mathrm{cs}}_{\pi^{\mathrm{cs}}_t} = \mathrm{FP32}\left[\mu_x + a_x\left(v_t + \frac{1}{n}\sum_{w=1}^{n}\left(\frac{s_{\pi^{\mathrm{cs}}_w} - \mu_x}{a_x} - v_w\right)\right)\right]. \tag{21}$$

Construction checks the stored sorting and the intended reference cuts. These scores encode supervision geometry; they are not inference-time inputs.

## C.2 REFINER AND STRUCTURAL SUPERVISION

**Inference-visible scoring.** For GAP window $k \in \{8, 16, 32, 64\}$, candidate features include frozen question and candidate representations, inference-visible ontology semantics and extraction

states, original score and rank, and bidirectional original-order score differences with validity masks:

$$\Delta^{-}_{i,d} = \frac{s_{i-d} - s_i}{a_x}, \quad \Delta^{+}_{i,d} = \frac{s_i - s_{i+d}}{a_x}, \quad d = 1, \ldots, k. \tag{22}$$

Out-of-range differences and their masks are zero. The base 205-dimensional representation and $4k$ GAP features enter a two-block local Transformer of width 128, with four attention heads and a 65-candidate attention window. A shared scalar head predicts one normalized residual per candidate:

$$s^{\theta}_i = s_i + a_x \delta_\theta(x)_i. \tag{23}$$

The forward pass uses neither reference evidence nor criterion-derived targets. All CLESO results reported in the main text use a separately trained GAP32 refiner for each domain at epoch 20 (GAP32/E20); CP and IP share the same refiner.

**Training objective.** Write $z_i = (s^{\theta}_i - \mu_x)/a_x$ and $z^{\mathrm{cs}}_i = (s^{\mathrm{cs}}_i - \mu_x)/a_x$. Role losses compare adjacent present priority groups within supported intervals. The sampled term averages higher–lower pairs; the hard term compares the lowest-scoring member of the higher group with the highest-scoring member of the lower group. Both apply $\mathrm{sp}_{0.02}(5{\times}10^{-4} - \Delta)$, where $\mathrm{sp}_T(u) = T\log(1 + \exp(u/T))$.

The endpoint outer-gap loss penalizes adjacent boundary gaps outside $[0.002, 0.004]$; its hard counterpart penalizes a difference below $0.002$ between the minimum score in one interval and the maximum score in the next. The sampled internal-gap loss penalizes adjacent drops outside $[0, 0.00075]$. The hard internal term penalizes the largest predicted adjacent gaps, up to 1,024, whose candidates share a reference interval identity. Sparse reference fitting applies Smooth-L1 with transition $0.05$ to $z_i - z^{\mathrm{cs}}_i$. Background drift applies Smooth-L1 with transition $0.1$ to $(s^{\theta}_i - s_i)/a_x$ at eligible background positions. These implemented penalties use explicit score bounds and sampled comparisons; they do not evaluate the Approach's ideal boundary inequality on every candidate pair.

With eligible comparisons averaged within each term, the per-question objective is

$$\begin{aligned}\mathcal{L}_{\mathrm{score}} = {} & 0.5\mathcal{L}_{\mathrm{role}} + \mathcal{L}_{\text{hard-role}} + 0.1\mathcal{L}_{\mathrm{out}} + 0.2\mathcal{L}_{\text{hard-out}} \\ & + 0.25\mathcal{L}_{\mathrm{in}} + \mathcal{L}_{\text{hard-in}} + 0.2\mathcal{L}_{\mathrm{ref}} + 0.05\mathcal{L}_{\mathrm{drift}}.\end{aligned} \tag{24}$$

Role comparison samples contain at most 32 pairs per adjacent present-role transition; ordinary internal-gap samples contain at most 1,024 pairs. Reference fitting includes role and boundary endpoints plus sampled positions up to a nominal 2,048; mandatory endpoints are retained. Drift samples at most 1,024 positions outside the completed-interval mask. That eligible set is empty whenever a criterion region exists, so the drift term is then zero. Empty loss terms contribute zero.

Training backpropagates $0.1$ times the mean per-question objective in each batch. AdamW runs for 20 epochs with learning rate $10^{-4}$, weight decay $10^{-4}$, four questions per batch, and gradient clipping at 1.

### C.3 Interval Recovery and Budgeted Observation

**Autonomous recovery.** Candidates are stably sorted by decreasing $s^{\theta}_i$, with original position breaking ties. The fixed decoder recursively examines normalized adjacent score gaps. A segment shorter than 256 candidates stops. Otherwise, legal cuts leave at least 128 candidates on each side; for a segment longer than 1,024, each side must also contain at least $\lceil m/4 \rceil$ of its $m$ candidates. The decoder chooses the largest legal gap, breaking ties by proximity to the midpoint and then by smaller cut index. Segments longer than 1,024 split even without a score drop; segments of length 256–1,024 split only when the winning normalized gap exceeds $0.001$. The rule recurses on both children. It produces $\widehat{\mathcal{I}}_x = (I_1, \ldots, I_{K_x})$ from predicted scores and fixed constraints, without accessing criterion intervals or target labels.

Reference-score construction and inference recovery are separate operations. The structural-transfer result in Appendix D explicitly states the conditions under which their interval memberships agree.

**Probe and local prefixes.** The separately trained, frozen domain Probe receives $q$, visible statistics of each recovered interval, and its affordable head prefix. These preliminary prefixes divide an independent budget $B_{\mathrm{probe}}$ equally across $K_x$ intervals. Statistics include candidate count, source-token cost, and refined-score summaries. For $K_x > 1$, the Probe's interval Yes–No logit differences

$\ell_j$ yield

$$\hat{r}_j = \frac{\sigma(\ell_j)}{\sum_{h=1}^{K_x} \sigma(\ell_h)}; \quad K_x = 1 \implies \hat{r}_1 = 1. \tag{25}$$

Its training target is each interval's fraction of annotated target occurrences. Questions without target occurrences provide no count-share target. The Probe is trained separately and held fixed during refiner training and evaluation.

For predicted local order $I_j = (p_{i_1}, \ldots, p_{i_m})$, let $\mathrm{Head}(I_j, b)$ be its longest complete prefix satisfying $\sum_{u=1}^{h} c_{i_u} \le b$. The main setting uses separate $B_{\mathrm{probe}} = B_{\mathrm{obs}} = 3250$ source-token budgets. Ordered concatenation gives

$$O = g(x) = \mathrm{Head}(I_1, B_{\mathrm{obs}}\hat{r}_1) \| \cdots \| \mathrm{Head}(I_{K_x}, B_{\mathrm{obs}}\hat{r}_{K_x}). \tag{26}$$

An unaffordable occurrence ends its interval's prefix; it is neither skipped nor truncated. Unused quota is not redistributed. Probe-read text enters $O$ only if an observation quota also admits it. Consequently, $\ell(O) \le B_{\mathrm{obs}}$, while probing incurs its own source-reading and model costs.

**Algorithm 1** Inference-time CLESO-CP

**Require:** $q, P, \mathbf{s}$, inference-visible ontology features, frozen refiner, Probe, reranker, and LLMs, $B_{\mathrm{probe}}, B_{\mathrm{obs}}$
**Ensure:** Selected evidence $\hat{E}$ and answer $\hat{a}$
1: Predict $s_i^{\theta} = s_i + a_x \delta_\theta(x)_i$ for every $p_i$
2: $\widehat{\mathcal{I}}_x \leftarrow \textsc{Recover}(\mathbf{s}^{\theta}, \pi_P, a_x)$
3: Read $\mathrm{Head}(I_j, B_{\mathrm{probe}}/K_x)$ for each recovered interval
4: Predict $\hat{\mathbf{r}}$ with the frozen Probe, or set $\hat{r}_1 = 1$ when $K_x = 1$
5: $O_j \leftarrow \mathrm{Head}(I_j, B_{\mathrm{obs}}\hat{r}_j)$; $O \leftarrow O_1 \| \cdots \| O_{K_x}$
6: $\tilde{a} \leftarrow L(q, O)$
7: $\mathcal{W}_x \leftarrow \textsc{PackCP}(q, \tilde{a}, (O_j)_{j=1}^{K_x})$ using Eq. 28
8: $\mathcal{W}_x^{\mathrm{sel}} \leftarrow \textsc{SelectLLM}(q, \tilde{a}, \mathcal{W}_x)$
9: $\hat{E} \leftarrow$ complete selected packages, restored to observation order
10: $\hat{a} \leftarrow L(q, \hat{E})$ in an independent context
11: **return** $\hat{E}, \hat{a}$

### C.4 Proposal-Conditioned Packages and Selection

The fixed answerer first generates $\tilde{a} = L(q, O)$. A frozen Qwen3-Reranker-4B scores each growing package $W$ using the question, fixed proposal, and complete package source text:

$$v(W) = \mathrm{logit}_{\mathrm{yes}}(q, \tilde{a}, W) - \mathrm{logit}_{\mathrm{no}}(q, \tilde{a}, W). \tag{27}$$

This is a packing proxy, not a probative probability. CLESO-IP grows packages separately within observed intervals. CLESO-CP visits the first occurrence of each nonempty interval, then the second, and so on. In CP, each local depth is a separate stream and its final package is flushed before the next depth. CP and IP share the recovered intervals and observation pool; they differ in package traversal.

For a current nonempty package $W$, an appended occurrence $p$ remains in that package when

$$v(W \| [p]) - v(W) \ge -(\tau_{\mathrm{drop}} + \tau_{\mathrm{num}}), \tag{28}$$

subject to fixed occurrence-, evidence-token-, and reranker-input caps. Otherwise $W$ is finalized and $p$ seeds the next package. A cap also closes the current package; an oversized singleton is truncated. Every successful traversal assigns each observed occurrence to exactly one nonempty package. The rule cuts at every score decrease only if both tolerances are zero.

The frozen LLM receives $(q, \tilde{a}, \mathcal{W}_x)$ and the source-bound contents of all packages in one request, then returns selected package IDs. A package is retained in full; when used in a set union, $W$ denotes all source-grounded spans carried by its occurrences. Selected evidence is restored to the occurrence order in $O$:

$$\hat{E} = f(q, O) = \bigcup_{W \in \mathcal{W}_x^{\mathrm{sel}}} W \preceq O. \tag{29}$$

The final answerer receives only $(q, \hat{E})$ in an independent context. Neither the proposal nor unselected text is carried into that call. Package scoring and selection can assemble joint support but do not certify $\hat{E} \in \mathcal{U}_{x,y}$.

### C.5 Information and Cost Boundaries

Criterion search and reference-score construction use train/dev annotations. Test-time feature construction, scoring, recovery, probing, packing, and selection use $q$, $P$, original scores, inference-visible ontology features, frozen models, and text read under the applicable budgets. LegalBench-RAG train/dev criterion construction uses an original-ranked prefix of at most one million source tokens, normalized using the full original sequence; test inference uses the full sequence. Benchmark and metric eligibility are specified in the experimental appendix.

$B_{\text{obs}}$ limits the source content of the selection pool, not total online computation. Online cost also includes full-sequence feature construction and refiner scoring, sorting and recovery, independent probing, proposal generation, repeated reranker scoring, LLM package selection, and final answering. Offline criterion construction and training are amortized rather than free. Evaluation reports these stages separately so that source-reading savings can be assessed against the added computation.

### C.6 Prompt for Greedy Package Scoring

Greedy packing evaluates a growing evidence package with a frozen binary reranker. The recorded interface places the question in `Query` and a compact JSON representation of the proposal-conditioned current set in `Document`. Its Yes–No logit difference provides the proxy score used to compare a package before and after an occurrence is appended. The shared input format and set-value instruction are reproduced below.

**Prompt C.1: Binary reranker input format**

```
<|im_start|>system
Judge whether the Document meets the requirements based on the
Query and the Instruct provided. Note that the answer can only be
"yes" or "no".<|im_end|>
<|im_start|>user
<Instruct>: {instruction}
<Query>: {question}
<Document>: {document}<|im_end|>
<|im_start|>assistant
<think>
</think>
```

**Prompt C.2: State-conditioned set-value instruction**

```
Assess the utility of the complete Current_Set for independently
answering the Query. The actual Answer_Proposal may be wrong or
incomplete: assess whether the source evidence supports, corrects,
or completes it. Consider redundancy, contradiction, and
complementary evidence jointly. Yes means higher set utility. Do
not assume unannotated material is false.
```

The score is a proxy for placing package boundaries, not a test of whether a package is probative. A separate frozen LLM subsequently selects a subset of the completed packages; the binary prompt above does not perform that selection.

# D Structural Transfer and the Cost of Evidence Access

This appendix proves conditional links between predicted scores, interval recovery, and the budget needed to expose a supporting witness. It then identifies the source-text saving available from local prefixes and accounts for downstream processing. These are deterministic results about access, not guarantees that training attains the stated conditions or that the final selector returns a probative set.

## D.1 Analysis Objects and Accounting

Retain $x = (q, P)$, $y = (E^*, a^*)$, $\mathcal{U}_{x,y}$, $O = g(x)$, and $\hat{E} = f(q, O) \preceq O$ from the Preliminary. Fix a nonempty witness $E^\dagger \in \mathcal{U}_{x,y}$ and a candidate cover $A^\dagger \subseteq \{1, \dots, n\}$ satisfying

$$\forall\, e \in E^\dagger,\ \exists\, i \in A^\dagger : e \subseteq_{\mathrm{char}} p_i. \tag{30}$$

The cover may choose one occurrence for a span appearing in several candidates. Witness and cover remain fixed throughout each comparison. Taking $E^\dagger = E^*$ invokes the witness assumption.

Write $\mathcal{I}_x^{\mathrm{cs}} = (I_1, \dots, I_K)$ for the criterion-search partition and $\mathbf{s}^{\mathrm{cs}}$ for its reference scores. Decreasing reference scores, with original-position tie-breaking, define the reference local and global orders. We identify the intervals of the recovered partition $\widehat{\mathcal{I}}_x$ with $I_1, \dots, I_K$ only after establishing equality of their ordered memberships.

Each occurrence $p_i$ has a positive charged source-token cost $c_i$. A prefix costs the sum of its occurrence charges, as in $\ell(O)$; overlapping occurrences remain separately charged. Separators, prompts, generated tokens, and repeated model computation are excluded from this source-content charge and included in the appropriate online ledger. Reference scores, interval memberships, and the fixed witness cover are supervision or hindsight quantities, not inference inputs.

## D.2 Conditions for Structural Recovery

Consider the fixed maximum-gap decoder under the count-determined legality, traversal, and stopping assumptions of Theorem 2. Suppose it returns $\mathcal{I}_x^{\mathrm{cs}}$ when run on $\mathbf{s}^{\mathrm{cs}}$. Define its normalized cross-interval separation

$$b_x = \min_{j<K} \frac{\min_{i \in I_j} s_i^{\mathrm{cs}} - \max_{i \in I_{j+1}} s_i^{\mathrm{cs}}}{a_x}. \tag{31}$$

Let $g_r^{\mathrm{cs}}$ denote the normalized adjacent score gap after sorted rank $r$. At reference decoding step $t$, let $J_t$ contain its legal cuts and $\rho_t$ be the selected cut. Its winning-gap margin is

$$\gamma_x = \min_{t:\, |J_t|>1} \left( g_{\rho_t}^{\mathrm{cs}} - \max_{r \in J_t \setminus \{\rho_t\}} g_r^{\mathrm{cs}} \right). \tag{32}$$

An empty minimum is $+\infty$; forced cuts need no gap comparison.

**Theorem 10** (Conditional transfer of interval memberships)**.** *Suppose the reference partition is recovered from* $\mathbf{s}^{\mathrm{cs}}$ *and* $b_x, \gamma_x > 0$*. If*

$$\frac{\|\mathbf{s}^\theta - \mathbf{s}^{\mathrm{cs}}\|_\infty}{a_x} \le \nu < \min\{b_x/2, \gamma_x/4\}, \tag{33}$$

*then decoding the predicted scores returns the same ordered interval memberships. Within-interval order may differ.*

*Proof.* The reference scores of adjacent intervals are separated by at least $a_x b_x$. Perturbing two scores changes their difference by at most $2a_x\nu$, so interval score bands remain strictly separated and preserve block membership after sorting.

If two score vectors differ coordinatewise by at most $a_x\nu$, their decreasing order statistics differ by at most $a_x\nu$ at every rank. Hence every normalized adjacent gap changes by at most $2\nu$, and the difference between a selected and competing gap changes by at most $4\nu$. The assumed winning margin keeps every selected legal cut strictly preferable. Induction over the decoder path preserves its cuts, legal sets, and count-determined stopping decisions. The resulting rank blocks agree, and band separation gives the same candidate memberships. □

The theorem requires both reference recovery and a positive decision margin. Boundary gaps exceeding neighboring internal gaps alone do not determine a recursive decoder's complete cut path. If a deployed decoder also uses score-threshold stopping, threshold decisions require their own stability margin before this result can be applied to that decoder.

### D.3 Ordering Error and Local Reading Cost

Let $A_j^\dagger = A^\dagger \cap I_j$. For nonempty $A_j^\dagger$, let $H_j^\dagger$ be the shortest reference-order prefix of $I_j$ containing all its required indices. Define

$$h_j = \sum_{k \in H_j^\dagger} c_k, \qquad H = \sum_{j=1}^{K} h_j. \tag{34}$$

For an empty intersection, let $H_j^\dagger = \varnothing$ and $h_j = 0$. On the same interval membership, let $h_j^\theta$ be the shortest prefix cost under the predicted local order that contains $A_j^\dagger$.

Write $k \prec_\theta i$ when $k$ precedes $i$ in the predicted stable order. The source-token cost of candidates outside the reference prefix that move ahead of any required candidate is

$$D_j^\theta = \sum_{k \in I_j \setminus H_j^\dagger} c_k \, \mathbf{1}\{\exists i \in A_j^\dagger : k \prec_\theta i\}. \tag{35}$$

**Lemma 11** (Ordering error and prefix cost)**.** *For every predicted local order,*

$$h_j^\theta \le h_j + D_j^\theta. \tag{36}$$

*Both sides are zero for an interval without required indices.*

*Proof.* Every occurrence in the required predicted prefix either belongs to $H_j^\dagger$ or is outside that reference prefix and precedes a required occurrence. The first group costs at most $h_j$ and the second is charged by $D_j^\theta$. □

This lemma assigns a source-token consequence to ordering errors. Moving a long unnecessary occurrence before required evidence costs more than rearranging occurrences already inside the necessary prefix. It does not require every witness candidate to have criterion-head supervision; uncovered candidates may simply have a larger reference demand $h_j$.

A prediction-error certificate is also possible. For $i \in I_j$, define

$$U_i(\nu) = \sum_{k \in I_j :\, s_k^{\mathrm{cs}} \ge s_i^{\mathrm{cs}} - 2a_x\nu} c_k. \tag{37}$$

Under Eq. 33, $k \prec_\theta i$ implies $s_k^{\mathrm{cs}} \ge s_i^{\mathrm{cs}} - 2a_x\nu$. Therefore, $h_j^\theta \le \max_{i \in A_j^\dagger} U_i(\nu)$, with an empty maximum equal to zero. This includes possible rearrangements among near-tied candidates.

The implemented role and gap losses encourage local priorities and boundary separation, respectively. Their sampled, unweighted comparisons do not directly upper-bound $D_j^\theta$ for every witness, particularly one containing uncovered targets.

### D.4 Allocation and Local-Prefix Savings

The actual quota threshold can be stated without reference to recovery. For each recovered interval, let $\widehat{h}_j$ be its shortest predicted-order prefix cost containing its indices from the fixed cover, or zero if it contains none. Hold scores, the separate probing budget, and predicted shares $\hat{r}_j$ fixed while varying $B_{\mathrm{obs}}$. The smallest real-valued charged observation budget that includes the cover is

$$B_{\mathrm{req}} = \max_{j:\, \widehat{h}_j > 0} \frac{\widehat{h}_j}{\hat{r}_j}. \tag{38}$$

A required interval with zero share gives $B_{\mathrm{req}} = +\infty$; an empty cover gives zero. Each independent quota includes its required prefix exactly when $B_{\mathrm{obs}}\hat{r}_j \ge \widehat{h}_j$. Integer-valued budgets require $\lceil B_{\mathrm{req}} \rceil$. No unused quota is redistributed.

When reference memberships are recovered, $\widehat{h}_j = h_j^\theta$. For $h_j > 0$, define the labeled-case deficiencies

$$\varepsilon_{\mathrm{ord}} = \max_{j:\, h_j>0} \frac{D_j^\theta}{h_j}, \qquad \varepsilon_{\mathrm{alloc}} = \max_{j:\, h_j>0} \left[1 - \frac{H\hat{r}_j}{h_j}\right]_+ . \tag{39}$$

They are diagnostic quantities, not inference inputs.

**Theorem 12** (Budget inflation for learned access)**.** *If reference memberships are recovered and every required interval has positive allocation, then* $\varepsilon_{\mathrm{alloc}} < 1$ *and*

$$B_{\mathrm{req}} \le \frac{1+\varepsilon_{\mathrm{ord}}}{1-\varepsilon_{\mathrm{alloc}}}\, H. \tag{40}$$

*Any* $B_{\mathrm{obs}}$ *at least this bound exposes* $E^\dagger$ *and makes* $\mathcal{U}_{x,y}(O)$ *nonempty.*

*Proof.* Lemma 11 gives $h_j^\theta \le (1+\varepsilon_{\mathrm{ord}})h_j$. The allocation deficiency gives $\hat{r}_j \ge (1-\varepsilon_{\mathrm{alloc}})h_j/H$. Their ratio is bounded by the right-hand side of Eq. 40 for every required interval. Taking the maximum and using Eq. 38 proves the budget claim. The resulting prefixes contain all cover occurrences, hence all spans of $E^\dagger$. □

The actual Probe learns shares of annotated target occurrences, not shares of prefix token demand. Let $r_j^{\mathrm{cnt}}$ denote its count-share target. When every required interval has positive $r_j^{\mathrm{cnt}}$, define $\kappa_x = \max_{j:h_j>0}(h_j/H)/r_j^{\mathrm{cnt}}$. If $\hat{r}_j \ge (1-\varepsilon_{\mathrm{cnt}})r_j^{\mathrm{cnt}}$ for required intervals, where $0 \le \varepsilon_{\mathrm{cnt}} < 1$, the same argument gives

$$B_{\mathrm{req}} \le \kappa_x \frac{1+\varepsilon_{\mathrm{ord}}}{1-\varepsilon_{\mathrm{cnt}}}\, H. \tag{41}$$

The factor $\kappa_x$ exposes the difference between target-count allocation and token demand. Perfect count-share prediction alone does not make allocation demand-optimal.

**Source-text saving relative to a global prefix.** Let $C_j = \sum_{i\in I_j} c_i$ and let $j_*$ be the last interval containing a cover index. In the same reference global order,

$$B_{\mathrm{glob}}^{\min} = \sum_{j<j_*} C_j + h_{j_*}, \qquad B_{\mathrm{glob}}^{\min} - H = \sum_{j<j_*} (C_j - h_j) \ge 0. \tag{42}$$

A global prefix reads all earlier intervals in full. Independent local prefixes skip their unused suffixes, including entire intervals without required evidence. Thus

$$\frac{1+\varepsilon_{\mathrm{ord}}}{1-\varepsilon_{\mathrm{alloc}}}\, H \le B_{\mathrm{obs}} < B_{\mathrm{glob}}^{\min} \tag{43}$$

is sufficient for the learned local policy to expose the fixed cover when that global prefix cannot. This does not rule out a different witness or duplicate cover in the global prefix. With one interval, the two prefix-access rules coincide in the same order. Comparison with the original relevance order requires its own fixed-cover global-prefix cost; Eq. 42 does not transfer unchanged between different orders.

### D.5 Probative Selection and Online Computation

Witness access gives the selector an opportunity, not a certificate of final success. For packages $\mathcal{W}_x$, let each $W$ denote its complete source-grounded spans. Whole-package selection can express the family

$$\mathcal{E}_x(\mathcal{W}_x) = \left\{ \bigcup_{W\in\mathcal{T}} W : \mathcal{T} \subseteq \mathcal{W}_x \right\}.$$

The success events obey

$$\{\hat{E} \in \mathcal{U}_{x,y}\} \subseteq \{\mathcal{E}_x(\mathcal{W}_x) \cap \mathcal{U}_{x,y} \neq \varnothing\} \subseteq \{\exists E \in \mathcal{U}_{x,y} : E \preceq O\}. \tag{44}$$

The first inclusion follows from whole-package selection; the second follows from source grounding. A useful span bundled with incompatible material can prevent package feasibility despite witness access, and the selector can fail even when a probative package union exists.

In a consistent runtime or compute-cost unit, the online ledger is

$$C_{\text{online}} = C_{\text{feat}} + C_{\text{ref}} + C_{\text{sort}} + C_{\text{dec}} + C_{\text{probe}} + C_{\text{proposal}} + C_{\text{pack}} + C_{\text{selector}} + C_{\text{answer}}. \tag{45}$$

Given constructed features and fixed network depth, width, and attention window, local Transformer scoring is $O(n)$ and stable sorting is $O(n \log n)$. Gap construction is $O(n)$; scanning legal cuts costs $O(\sum_t |J_t|)$ and can be quadratic under repeated unbalanced scans. Criterion construction is offline, but a forward scan alone does not make repeated criterion search linear.

**Proposition 13** (Downstream source-processing bound)**.** *Let $m_O$ be the number of observed occurrences. For one append-once greedy packing traversal with at most one singleton restart per rejected extension, no retries, and one proposal, selector, and final-answer request,*

$$M_{\text{rank}} \le \max\{0, 2m_O - 1\}. \tag{46}$$

*In a common source-token accounting unit, downstream calls process at most*

$$T_{\text{down,src}} \le (M_{\text{rank}} + 3)\ell(O) \le (2m_O + 2)B_{\text{obs}}. \tag{47}$$

*If every observed occurrence costs at least $c_{\min} > 0$, then $m_O \le \lfloor B_{\text{obs}}/c_{\min} \rfloor$.*

*Proof.* For $m_O \ge 1$, introducing an occurrence requires at most one initial or appended-package score. Each of at most $m_O - 1$ rejected extensions adds one singleton restart score. This proves the first bound; an empty observation makes no package request.

Every scored package draws from $O$ and contains at most $\ell(O)$ charged source text. The proposal reads $O$, the selector reads the nonoverlapping packages, and the final answerer reads selected source text; these contribute at most three more copies of $\ell(O)$. Substitution of the request and observation bounds gives Eq. 47. Finally, $m_O c_{\min} \le \ell(O) \le B_{\text{obs}}$. □

The proposition applies to each fixed IP or CP traversal separately. Independent probing adds at most $B_{\text{probe}}$ charged source tokens. Questions, interval metadata, proposal text repeated in reranker requests, generated tokens, and full-sequence features remain outside this source-only bound but inside the online ledger. Different model tokenizers must be reported in their own units.

Repeated package scoring can remain costly. If $m_O$ equally sized occurrences form one growing package and $B_{\text{obs}} = m_O c$, scoring every prefix reads $c \sum_{t=1}^{m_O} t = (m_O + 1)B_{\text{obs}}/2$ source tokens. Caching can reduce executed work; retries increase it and must be charged.

These results give a conditional access mechanism: recoverable boundaries and sufficiently accurate local priorities and allocation expose a fixed witness at a budget governed by local prefix costs. They bound downstream source-text exposure, not total runtime or final Probative success. Net computational benefit requires the complete measured cost ledger.

# E EXPERIMENTAL DETAILS

This appendix specifies the comparison protocol behind Table 2 and reports diagnostics of structural supervision, budgeted observation, and online cost.

## E.1 DETAILS OF DATA, OBSERVATION BUDGETS, COMPARISON METHODS, AND METRICS

**Evaluation scope.** The three benchmarks differ in where supporting evidence must be found. BioASQ links biomedical questions to snippets from PubMed article titles and abstracts (Tsatsaronis et al., 2015). QASPER poses questions based on an NLP paper's title and abstract but requires answers from its full text, often supported by multiple paragraphs (Dasigi et al., 2021). LegalBench-RAG anchors legal questions to precise text spans in agreements and privacy policies, rather than whole-document matches (Pipitone & Alami, 2024). Our processed Train/Dev/Test splits contain 2,439/264/260, 2,593/1,005/1,451, and 4,941/536/1,412 questions, respectively. Complete Test candidate sequences average 885,449, 4,904, and 2,088,524 source tokens per question, in the same order. LegalBench-RAG Train and Dev use a RAW-ranked prefix capped at one million source tokens, whereas Test inference processes the complete candidate sequence. These lengths sum the costs of all candidate occurrences; they are not single LLM context lengths.

**Ontology resources.** The resources used to construct $M_x$ are fixed across splits. BioASQ uses a local SQLite build of UMLS 2026AA for biomedical concepts, semantic types, and hierarchical relations; MeSH is not used as a separate ontology (U.S. National Library of Medicine, 2026). QASPER uses CSO v3.5 topics and direct-parent relations, together with eight DEO discourse-section categories (Computer Science Ontology, 2025; Shotton & Peroni, 2026). LegalBench-RAG uses five facets of noslegal v4 (local source file v4.02): work types, areas of law, participants, roles, and information assets. It additionally uses only the Core and Personal Data modules of DPV v2.3 (noslegal, 2026; Data Privacy Vocabularies and Controls Community Group, 2026). Inference-time ontology features are derived from the question, candidates, and these fixed resources; evidence labels and fitted criteria are used for training supervision, not refiner input.

**Evidence and budgets.** Candidate occurrences retain their source identities and offsets. BioASQ and LegalBench-RAG evidence is evaluated by frozen source-character alignment; QASPER uses its frozen paragraph-evidence mapping. An occurrence incurs its complete source-token cost even if source text overlaps another occurrence. The comparison applies a common 3,250-source-token cap to final selected evidence. For CLESO, $B_{\text{obs}} = 3{,}250$ additionally bounds the source text admitted to $O$, and an independent $B_{\text{probe}} = 3{,}250$ probes interval heads. Complete occurrence costs are charged even when source text overlaps. The observation cap does not represent feature processing, probe input, repeated package scoring, or answer-generation cost; these quantities are accounted for separately.

**Compared Methods** The comparison includes the original relevance prefix, BM25 (Robertson & Zaragoza, 2009), BGE and Qwen3 rerankers (Zhang et al., 2025), and BioASQ-only MedCPT (Jin et al., 2023). MMR (Goldstein & Carbonell, 1998), Adaptive-$k$, RECOMP-extractive (Xu et al., 2024), and Provence-pretrained (Chirkova et al., 2025) vary redundancy control or compression. RankGPT-Qwen (Sun et al., 2023), SetR-prompt-Qwen (Lee et al., 2025), BudgetedCoverage, PACE-Frontloading (Cai & Zafarani, 2026), AdaGReS (Peng et al., 2025), METEORA-Qwen (Saxena et al., 2025), and IRCoT-Qwen (Trivedi et al., 2023) provide listwise, setwise, or iterative alternatives. SetR uses prompting without domain-specific fine-tuning; Provence uses pretrained weights. PACE uses fixed-budget SPLADE-v3 frontloading rather than its pressure-adaptive controller, and METEORA uses the base model without its Verifier. Each row is evaluated from its saved final evidence set; a common evaluator does not imply identical internal selectors. The separate budget replay compares six baselines: BM25, BGE, Qwen3-Reranker-4B, Provence, SetR, and METEORA.

**Metrics and Valid Outcomes** Evidence precision, recall, and F1 evaluate the final selected set $\hat{E}$ using each benchmark's source mapping. Values are question-level scores macro-averaged over defined cases, so macro F1 is not the harmonic mean of macro precision and recall. QASPER Answer is official answer-token F1; BioASQ and LegalBench-RAG Answer are frozen model-assisted answer judgments. Tail Recall measures selected evidence beyond a fixed original-prefix reference

Table 5: Full-Test Evidence F1 (%) under five observation budgets. The best baseline is the highest reported mean among the six budget-replay baselines.

| Benchmark | $B_{\text{obs}}$ | CLESO-CP | CLESO-IP | Best baseline F1 | Baseline |
|---|---|---|---|---|---|
| **BioASQ** | 1,000 | **22.90** | 20.60 | 21.65 | Qwen3-Reranker |
| | 2,000 | **24.35** | 22.71 | 22.85 | BGE-Reranker |
| | 3,250 | **26.95** | 24.33 | 26.40 | BGE-Reranker |
| | 5,000 | **26.02** | 22.59 | 23.47 | BGE-Reranker |
| | 10,000 | **24.94** | 20.57 | 23.49 | SetR |
| **QASPER** | 1,000 | **55.60** | 35.80 | 52.94 | SetR |
| | 2,000 | **55.85** | 35.29 | 53.17 | Qwen3-Reranker |
| | 3,250 | **56.25** | 33.73 | 54.19 | SetR |
| | 5,000 | **56.98** | 34.54 | 54.23 | BGE / Qwen3-Reranker |
| | 10,000 | **57.12** | 34.91 | 54.39 | BM25 / SetR |
| **LegalBench-RAG** | 1,000 | **17.25** | 12.51 | 16.37 | Qwen3-Reranker |
| | 2,000 | **20.04** | 14.11 | 18.45 | Qwen3-Reranker |
| | 3,250 | **29.98** | 19.03 | 26.25 | Qwen3-Reranker |
| | 5,000 | **22.83** | 15.27 | 21.63 | Qwen3-Reranker |
| | 10,000 | **24.30** | 16.22 | 23.04 | Qwen3-Reranker |

and is undefined when a question has no tail opportunity. The Probative (PB) metric is a reference-conditioned judgment of whether the selected evidence supports the answer requirements. It is an evaluation proxy, not certification that $\mathrm{Rel}_{x,y}(\hat{E}) = 1$. Saved outputs and references are held fixed during evaluation. Technical retries address malformed or truncated requests; a valid negative judgment is not retried to obtain a different decision. NA and UNKNOWN are not imputed as zero. PB eligibility in particular depends on qualified references, so its denominator cannot be replaced by the Evidence F1 denominator. Confidence intervals, when reported, resample questions with their paired method outputs; they do not measure variation across training seeds.

### E.2 Budgeted Observation and Component Interactions

The full-Test ablation fixes GAP32/E20 and the 4B package scorer. Before package selection, Probe observation reaches target recall of 34.78%, 94.51%, and 41.49% on BioASQ, QASPER, and LegalBench-RAG. With CP fixed, Probe rather than equal interval allocation raises final evidence F1 by 17.87, 2.11, and 26.58 points. Given the same Probe pool, CP rather than IP raises precision by 9.67, 23.90, and 14.73 points, while recall changes by only +0.43, +0.35, and -0.19 points. Allocation thus improves access to targets, whereas cross-interval packing chiefly improves final-set precision.

### E.3 Observation-Budget Sensitivity

Table 5 reports full-Test Evidence F1 at five observation budgets. Its best-baseline column takes the highest reported mean among the six methods in the budget replay, whereas the full main comparison includes additional baselines. CLESO-CP leads all 15 benchmark–budget cells in this replay. The five CLESO points use the same GAP32/E20 refiner and 4B package scorer. BioASQ and LegalBench-RAG reach their highest reported CP F1 at 3,250 tokens, while QASPER reaches 57.12% at 10,000 tokens. A larger observation pool can change package formation and final selection, so the selected evidence sets need not be nested. Only Evidence F1 has a complete five-budget replay.

### E.4 Recorded Online Cost

Table 6 distinguishes method-native evidence-selection cost from complete online cost on the full Test splits. The observation cap limits source text admitted to $O$, whereas model-input totals accumulate tokens across recorded calls. Times are serial-equivalent; feature construction, model loading, offline training, and evaluation judges are excluded.

Table 6: Recorded CLESO cost per question on the full Test splits at $B_{\text{obs}} = 3{,}250$ with a 4B package reranker.

| | | Method-native | | Complete online | |
|---|---|---|---|---|---|
| Benchmark | Variant | Time (s/q) | Input (k tok/q) | Time (s/q) | Input (k tok/q) |
| BioASQ | CP | 29.73 | 71.4 | 76.37 | 87.7 |
| | IP | 72.86 | 83.3 | 99.01 | 93.1 |
| QASPER | CP | 25.54 | 41.9 | 54.30 | 61.3 |
| | IP | 93.15 | 76.9 | 107.45 | 87.7 |
| LegalBench-RAG | CP | 16.59 | 53.6 | 46.92 | 68.3 |
| | IP | 52.94 | 64.2 | 67.50 | 72.5 |

Method-native costs match Table 3. Complete-online costs additionally include independently generated proposals, package selection, and final answering. Model input is cumulative across recorded calls, not the source-text observation size.

### E.5 LIMITATIONS

The main comparison, ablation, observation statistics, cost analysis, and budget replay use the full BioASQ, QASPER, and LegalBench-RAG Test splits. Reference-conditioned PB is a model-assisted support measure, not certification of the ideal Probative condition; we measure the answer differently across the three benchmarks.

Criterion construction has access to training labels, but the deployed refiner does not. The constructed regions cover 56.5% of QASPER and 83.2% of LegalBench-RAG Train target occurrences; uncovered targets receive no criterion-based head promotion. Refiner Dev global-prefix recall remains close to the original order and below the criterion sequence, especially on BioASQ and LegalBench-RAG. The empirical advantage of CP therefore concerns the combined local-observation and selection policy, not faithful recovery of every label-visible interval or a general improvement in global ranking.

The 3,250-token limit applies to source text admitted to $O$, not to the independent Probe, full-candidate feature processing, repeated pack scoring, or answer generation. Recorded full-chain costs omit feature-construction time, and corresponding baseline stages have not been measured under the same boundary. The five-budget replay holds the 4B package scorer fixed. BioASQ and LegalBench-RAG peak at 3,250 tokens among the tested budgets, whereas QASPER reaches its highest reported CP F1 at 10,000 tokens; these observations do not establish a universal optimal budget.