REPETITION OR CORROBORATION? EVIDENCE SELECTION FOR RETRIEVAL-AUGMENTED GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Retrieval-augmented generation (RAG) ranks passages by relevance, but agreement among passages does not guarantee independent evidential support. Several passages may repeat one evidence basis, causing redundant support and discarding useful corroboration from distinct sources. We introduce BasisDPP-RAG, which separates the query-conditioned claim from its evidence basis and combines them in a factorized, quality-aware determinantal point process (DPP) for budgeted evidence selection. The resulting objective penalizes same-basis repetition while preserving agreement across distinct bases. Experiments on AVeriTeC, NLI4CT, and SciFact yield 84.84% average Exact Match, 11.17 percentage points above the strongest baseline, while reducing NLI4CT basis redundancy from 57.76% to 7.89% and improving robustness to adversarial distractors. These results show that reliable RAG should measure evidential support by distinct bases rather than passage count.

1 INTRODUCTION

Retrieval-augmented generation (RAG) supplements language models' incomplete or outdated knowledge with retrieved external text (Lewis et al., 2020). Dense retrieval improves access to relevant passages (Karpukhin et al., 2020), while self-reflection and corrective retrieval help assess and filter retrieved context (Asai et al., 2024; Yan et al., 2024). However, passage-level relevance does not establish whether retrieved passages provide independent evidential support.

This distinction matters when retrieved passages share an underlying evidence basis. A trial report , press release, news articles, and database summary may all restate one study. A relevance-based selector can therefore fill the context with highly relevant passages while providing only one underlying basis of support. Conversely, semantic deduplication may remove an independent study that reaches the same conclusion. More relevant passages do not necessarily provide more independent support. Figure 1 shows how repeated passages can crowd out necessary evidence: relevance-based selection covers only Trial A, whereas BasisDPP-RAG covers both trials needed to verify the claim.

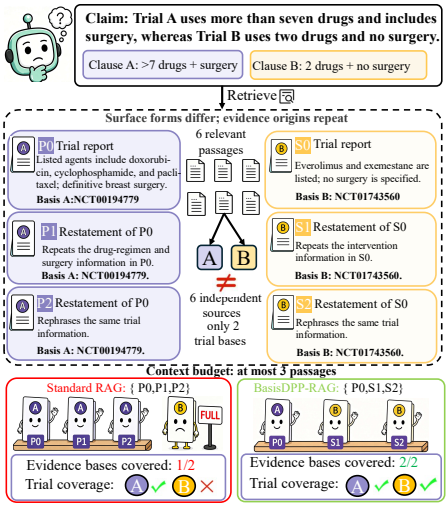


Figure 1: **Repeated passages can crowd out necessary evidence.** Verifying this claim requires evidence about both trials. Standard RAG fills all three slots with Trial A passages, whereas BasisDPP-RAG covers both trials.

Prior work shows that irrelevant context can distract language models and that relevant information may be underused in long inputs (Shi et al., 2023; Yoran et al., 2024; Liu et al., 2024). Repeated passages, however, may all be relevant and factually correct, so relevance filtering alone cannot identify their shared evidence basis. Truth-discovery research likewise shows that copying among sources can inflate apparent agreement (Dong et al., 2009).

Diversification methods such as maximal marginal relevance and determinantal point processes (DPPs) balance relevance against redundancy using a chosen similarity representation (Carbonell & Goldstein, 1998; Kulesza & Taskar, 2012). Lexical or semantic similarity may miss repeated

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.

accounts of one evidence basis or penalize corroboration across distinct bases. Evidence selection therefore requires modeling both the query-conditioned claim and its evidence basis.

We introduce **BasisDPP-RAG**, a trainable evidence selector for RAG. It separately encodes each passage's query-conditioned claim and evidence basis, then combines their similarities in a factorized kernel. A quality-aware DPP uses this kernel to select passages under cardinality and token constraints. By penalizing similarity jointly in the claim and basis views, the selector discourages repeated support while retaining independently grounded corroboration.

Our contributions are threefold:

- We formulate evidence selection under shared evidence bases, distinguishing query-conditioned claim agreement from evidence-basis dependence.
- We propose BasisDPP-RAG, a trainable selector that combines a factorized dependence kernel with quality-aware DPP selection under a context budget.
- We establish kernel validity and characterize the separation of repetition and corroboration under the modeled similarities. Experiments show improved accuracy across three benchmarks, with lower basis redundancy and greater distractor robustness on NLI4CT.

2 RELATED WORK

Retrieval-augmented language models. Early retrieval-augmented systems integrate external text into language-model pretraining and generation (Guu et al., 2020; Lewis et al., 2020; Borgeaud et al., 2022), while dense retrieval improves evidence retrieval for open-domain question answering (Karpukhin et al., 2020). More recent methods control when and how retrieved evidence is used. FLARE retrieves during generation (Jiang et al., 2023), Self-RAG uses reflection tokens to guide retrieval and critique (Asai et al., 2024), and CRAG evaluates retrieval quality to trigger corrective actions (Yan et al., 2024). Speculative RAG generates answer drafts in parallel (Wang et al., 2025), while CF-RAG incorporates counterfactual reasoning (Qin et al., 2026).

Retrieval robustness and context optimization. Retrieved context is not uniformly beneficial Irrelevant passages can distract language models (Shi et al., 2023; Yoran et al., 2024), and relevant information may be underused in long inputs (Liu et al., 2024). Methods for improving context utility include RECOMP, which learns compression with selective augmentation (Xu et al., 2024); FIT-RAG, which combines factual-information scoring with token reduction (Mao et al., 2025); and augmentation-adapted retrieval, which learns language models' document preferences (Yu et al., 2023). These methods improve the usefulness or efficiency of retrieved context. Evidence dependence presents an additional challenge: several passages may each be useful and factually correct while repeating the same underlying evidence.

Evidence retrieval for fact verification. Fact-verification benchmarks require selecting evidence to support a verdict. FEVER provides sentence-level evidence for claim verification (Thorne et al., 2018), while SciFact focuses on scientific claims and research abstracts (Wadden et al., 2020). AVeriTeC and NLI4CT extend this setting to heterogeneous web evidence and clinical-trial reports, respectively (Schlichtkrull et al., 2023; Jullien et al., 2023). Relevance and entailment help identify useful passages and assess their relation to a claim. Evaluating dependence among those passages adds a complementary perspective on the selected evidence set.

Diversity and source-dependent aggregation. Maximal marginal relevance balances relevance against redundancy through greedy selection (Carbonell & Goldstein, 1998), whereas DPPs score subsets jointly through a quality–diversity kernel (Kulesza & Taskar, 2012). Both rely on a similarity representation whose meaning depends on the application. Separately, truth-discovery research estimates source reliability (Yin et al., 2008) and models copying when aggregating conflicting information Dong et al. (2009). These approaches typically operate on structured claims and source identities. BasisDPP-RAG instead selects query-conditioned passages under a context budget. It combines claim agreement and evidence-basis dependence in a factorized kernel, using a quality-aware DPP to balance passage relevance with nonredundant support.

3 PRELIMINARIES

Given a query $q \in \mathcal{Q}$, corpus $\mathcal{D} = \{d_1, \ldots, d_N\}$, retriever R, and item-wise relevance score $s_q(e)$, standard RAG selects the top-k retrieved passages and generates:

$$\widehat{y}_{\text{Top-}k} = \arg\max_{y \in \mathcal{Y}} p_\phi(y \mid q, \text{TopK}(R(q, \mathcal{D}), s_q, k)) \,. \tag{1}$$

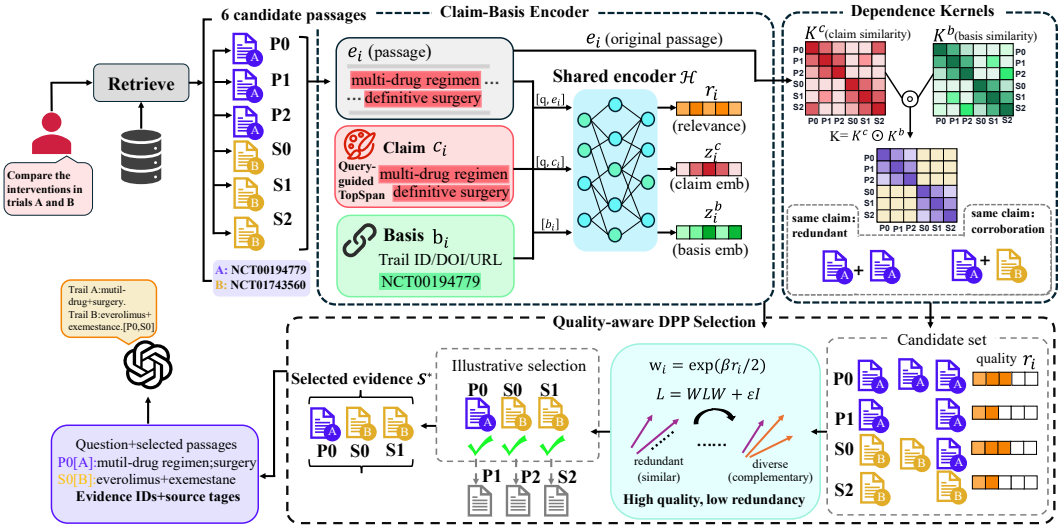


Figure 2: **Overview of BasisDPP-RAG.** Retrieved candidates are represented through their original passages, query-conditioned claims, and evidence bases. The claim and basis kernels capture complementary candidate relationships; together with item relevance, they guide quality-aware DPP selection. The selected evidence is then serialized for generation.

Here, p_ϕ is the generator's conditional distribution over the answer space $\mathcal{Y}$, and TopK returns the k highest-scoring candidate passages. This selection rule scores passages independently and does not account for shared evidence bases. Several highly ranked passages may therefore repeat the same underlying evidence.

Definition 1 (**k-Determinantal Point Process).** For a finite candidate set $\mathcal{E}$ and a positive-semidefinite kernel $L \succeq 0$, a k-DPP assigns each subset $S \subseteq \mathcal{E}$ with $|S| = k$ probability

$$p_L(S \mid |S| = k) = \frac{\det(L_S)}{Z_k(L)}, \qquad Z_k(L) = \sum_{|T|=k} \det(L_T), \tag{2}$$

provided $Z_k(L) > 0$. Here, L_S is the principal submatrix indexed by S, $\mathcal{E}$ is the retrieved candidate pool. The diagonal of L represents item quality, while its off-diagonal entries represent pairwise similarity, so the determinant favors high-quality, nonredundant subsets. However, a generic DPP cannot determine whether similarity reflects shared provenance or independent corroboration. BasisDPP-RAG resolves this ambiguity by factorizing the kernel into query-conditioned claim similarity and evidence-basis dependence.

4 METHODOLOGY

Retrieved passages may express the same query-relevant claim while either repeating one evidence basis or providing corroboration from distinct bases. BasisDPP-RAG distinguishes these cases by representing each candidate through its query-conditioned claim and evidence basis, then applying quality-aware determinantal selection to construct the evidence set passed to the generator. Figure 2 illustrates the overall architecture.

4.1 Problem Formulation and Overview

Given a query q and corpus $\mathcal{D}$, a first-stage retriever constructs a high-recall candidate pool:

$$\mathcal{E}(q) = \{e_1, \ldots, e_n\} = \text{Retrieve}(q, \mathcal{D}). \tag{3}$$

The selector chooses a compact subset $S^\star \subseteq \mathcal{E}(q)$ that balances query relevance with coverage of distinct resolved evidence bases, and the generator produces an answer using only the selected evidence:

$$S^\star = \text{Select}(q, \mathcal{E}(q)), \qquad \hat{y} = G_\phi(q, S^\star). \tag{4}$$

Equation equation 4 defines the retrieval–selection–generation interface. Given the retrieved candidate pool $\mathcal{E}(q)$, Select combines item relevance, query-conditioned claim agreement, and evidence-basis dependence to return a budget-feasible subset $S^\star$. These signals favor relevant evidence while discouraging repeated support that expresses the same claim and depends on the same resolved basis.

4.2 Query-Conditioned Claim–Basis Representation

Each retrieved candidate is represented by its original passage, query-conditioned claim, and evidence basis:

$$x_i = (e_i, c_i, b_i), \qquad b_i = (g_i, u_i) = \mathrm{ResolveBasis}(e_i, m_i; \tau_{\mathrm{basis}}), \tag{5}$$

where e_i is the original passage, c_i is the claim most relevant to query q, m_i denotes available provenance metadata, g_i is a canonical basis key, and u_i is a normalized readable basis descriptor. The passage is used for relevance estimation and generation, whereas the claim and basis views define pairwise relations.

Query-conditioned claim. Because a passage may contain multiple statements, we extract the span that most directly addresses the query. Let $e_i = (s_{i1}, \ldots, s_{im_i})$ denote its ordered sentence sequence. A fixed pretrained query–sentence scorer assigns

$$a_{it} = s_{\mathrm{sent}}(q, s_{it}), \qquad c_i = \mathrm{TopSpan}(\{s_{it}\}, \{a_{it}\}). \tag{6}$$

$\mathrm{TopSpan}$ retains the highest-scoring sentence and may include one adjacent sentence needed to preserve a reference, negation, quantity, or qualification. The sentence scorer is fixed during selector training; $\mathcal{L}_{\mathrm{claim}}$ supervises the representation of the extracted claim rather than the discrete span-selection operation.

Evidence basis. The claim view captures what a passage says, whereas the basis view captures where its information originates. For resolved provenance, g_i is a canonical identifier such as a DOI, trial identifier, report number, canonical URL, or normalized bibliographic tuple, and u_i is its corresponding readable descriptor. If provenance cannot be resolved, we set

$$g_i = \texttt{UNRESOLVED:ID}_i, \qquad u_i = \texttt{[UNK_BASIS]}, \tag{7}$$

so that unresolved passages remain distinct for bookkeeping but are not treated as independently identifiable sources by the basis encoder or basis-pair supervision.
A relevance head estimates passage quality:

$$r_i = \sigma\big(\mathbf{w}_r^\top H_\theta([q; e_i]) + d_r\big), \tag{8}$$

where $r_i \in (0, 1)$ measures how directly e_i supports answering q. The claim and basis heads produce normalized representations:

$$\mathbf{z}_i^c = \mathrm{Norm}(W_c H_\theta([q; c_i])), \tag{9}$$

$$\mathbf{z}_i^b = \mathrm{Norm}(W_b H_\theta(\tilde{u}_i)), \tag{10}$$

where $\tilde{u}_i = u_i$ for resolved bases and $\tilde{u}_i = \texttt{[UNK_BASIS]}$ otherwise. The claim representation is query-conditioned, whereas the basis representation excludes q so that provenance remains invariant to query wording. Benchmark-specific resolution and supervision details are given in Appendix B.3.

4.3 Factorized Evidence-Dependence Kernel

Neither claim similarity nor basis similarity alone suffices to identify redundant support: claim-only similarity may suppress cross-basis corroboration, whereas basis-only similarity may suppress complementary claims from the same basis. We therefore define separate RBF kernels for claim and basis representations:

$$K_{ij}^c = \exp\left(-\frac{\|\mathbf{z}_i^c - \mathbf{z}_j^c\|_2^2}{2\sigma_c^2}\right), \tag{11}$$

$$K_{ij}^b = \exp\left(-\frac{\|\mathbf{z}_i^b - \mathbf{z}_j^b\|_2^2}{2\sigma_b^2}\right). \tag{12}$$

Here, $\sigma_c > 0$ and $\sigma_b > 0$ control how rapidly similarity decreases with representation distance. We combine the two kernels through their Hadamard product:

$$K = K^c \odot K^b, \qquad K_{ij} = K_{ij}^c K_{ij}^b. \tag{13}$$

Because K_{ij} is large only when both claim and basis similarities are high, the factorization penalizes passages that express the same claim and are assigned high similarity by the learned basis representation. Similar claims from distinct resolved bases remain jointly selectable when their learned basis similarity is low.

4.4 Quality-Aware Determinantal Selection

The factorized kernel K captures pairwise evidence dependence; we incorporate item quality through an L-ensemble DPP by converting the relevance score into a positive quality weight:

$$w_i = \exp\left(\frac{\beta r_i}{2}\right), \qquad W = \mathrm{diag}(w_1, \ldots, w_n), \qquad L = WKW + \varepsilon I. \tag{14}$$

Here, $\beta \geq 0$ controls the emphasis on item relevance and εI provides numerical stabilization. For the unregularized kernel $L^{(0)} = WKW$ and any subset S with $\det(K_S) > 0$, the subset objective decomposes as

$$\log\det\left(L_S^{(0)}\right) = \beta \sum_{i \in S} r_i + \log\det(K_S). \tag{15}$$

The first term rewards query-relevant passages, while the second favors basis-aware diversity; larger β places more weight on relevance. We use the conditional k-DPP defined in Equation 2 for fixed-size supervision during training. At inference, we allow fewer than k passages to satisfy the token budget and stop when no positive marginal gain remains. We therefore seek a budget-feasible subset:

$$S^\star = \arg\max_{S \subseteq \mathcal{E}(q)} \log\det(L_S), \qquad |S| \leq k, \qquad \sum_{i \in S} \ell_i \leq B. \tag{16}$$

Here, ℓ_i denotes the token length of the serialized evidence block for e_i, and B is the context budget. We approximate this MAP objective with greedy selection and incremental Cholesky updates, adding the feasible candidate with the largest marginal log-determinant gain. Selection stops when k passages have been selected, no candidate fits the remaining budget, or no positive gain remains.

4.5 Learning Objective

We train the selector with item-, pair-, and set-level supervision. For passage relevance, $y_i \in \{0, 1\}$ indicates whether e_i provides evidence useful for answering q, and the relevance loss is

$$\mathcal{L}_{\mathrm{rel}} = -\frac{1}{n} \sum_{i=1}^{n} \left[y_i \log r_i + (1 - y_i) \log(1 - r_i) \right]. \tag{17}$$

For each candidate pair (i, j), y_{ij}^c indicates whether the passages express the same query-relevant claim, while y_{ij}^b indicates whether they share an evidence basis. We predict these relations from the corresponding representations:

$$p_{ij}^c = \sigma\left(\tau_c (\mathbf{z}_i^c)^\top \mathbf{z}_j^c + \delta_c\right), \qquad \mathcal{L}_{\mathrm{claim}} = \mathrm{BCE}(p_{ij}^c, y_{ij}^c), \tag{18}$$

$$p_{ij}^b = \sigma\left(\tau_b (\mathbf{z}_i^b)^\top \mathbf{z}_j^b + \delta_b\right), \qquad \mathcal{L}_{\mathrm{basis}} = \mathrm{BCE}(p_{ij}^b, y_{ij}^b). \tag{19}$$

Here, τ_c and τ_b are learnable scales, δ_c and δ_b are biases, and BCE denotes binary cross-entropy. Distinct-basis corroboration corresponds to $y_{ij}^c = 1$ and $y_{ij}^b = 0$, while pairs with unresolved basis relations are masked from $\mathcal{L}_{\mathrm{basis}}$.

For set-level supervision, let $\mathcal{G}_k$ denote the annotated valid evidence subsets of size k. The set-level loss is

$$\mathcal{L}_{\mathrm{set}} = -\log \frac{\sum_{S \in \mathcal{G}_k} \det(L_S)}{Z_k(L)}, \tag{20}$$

which assigns probability mass to all annotated valid subsets. The complete objective is

$$\mathcal{L} = \lambda_{\mathrm{rel}} \mathcal{L}_{\mathrm{rel}} + \lambda_{\mathrm{claim}} \mathcal{L}_{\mathrm{claim}} + \lambda_{\mathrm{basis}} \mathcal{L}_{\mathrm{basis}} + \lambda_{\mathrm{set}} \mathcal{L}_{\mathrm{set}}. \tag{21}$$

The coefficients $\lambda_{\mathrm{rel}}, \lambda_{\mathrm{claim}}, \lambda_{\mathrm{basis}}, \lambda_{\mathrm{set}}$ balance the four terms. We first warm up the relevance and pairwise heads, then activate $\mathcal{L}_{\mathrm{set}}$ for joint training. Because the conditional k-DPP likelihood is differentiable, this procedure requires neither reinforcement learning nor a straight-through estimator.

Inference and generation. At inference, the system retrieves a high-recall candidate pool, constructs the query-conditioned claim and evidence-basis views, computes relevance and factorized dependence scores, and selects a budget-feasible subset using greedy DPP MAP inference. The selected evidence is serialized as

$$\begin{aligned} C(q, S^\star) &= [q; \{\mathrm{ID}_i, b_i, e_i\}_{e_i \in S^\star}; \text{attribution}], \\ \hat{y} &= G_\phi(C(q, S^\star)). \end{aligned} \tag{22}$$

Here, $C(q, S^\star)$ contains the query, selected passages, basis descriptors, and citation identifiers; the exact serialization and output protocol are given in Appendix F. For n candidates and representation dimension d, kernel construction costs $O(n^2 d)$, while greedy selection costs approximately $O(nk^2)$.

Algorithm 1 Inference of BasisDPP-RAG

Input: Query q, corpus $\mathcal{D}$, retriever R, selector F_θ, generator G_ϕ, cardinality k, and token budget B
Output: Answer $\hat{y}$ and selected evidence set $S^\star$

$\mathcal{E}(q) \leftarrow \mathrm{Retrieve}(q, \mathcal{D})$, with $|\mathcal{E}(q)| = n$
for $e_i \in \mathcal{E}(q)$ **do**
  $c_i \leftarrow \mathrm{TopSpan}(e_i; q)$
  $b_i \leftarrow \mathrm{ResolveBasis}(e_i, d_i; \tau_{\mathrm{basis}})$
  $(r_i, \mathbf{z}_i^c, \mathbf{z}_i^b) \leftarrow F_\theta(q, e_i, c_i, b_i)$
end for
$K_{ij}^c \leftarrow \exp\big(-\|\mathbf{z}_i^c - \mathbf{z}_j^c\|_2^2/(2\sigma_c^2)\big)$, for all i, j
$K_{ij}^b \leftarrow \exp\big(-\|\mathbf{z}_i^b - \mathbf{z}_j^b\|_2^2/(2\sigma_b^2)\big)$, for all i, j
$w_i \leftarrow \exp(\beta r_i/2)$
$W \leftarrow \mathrm{diag}(w_1, \ldots, w_n)$
$L \leftarrow W(K^c \odot K^b)W + \varepsilon I$
$S^\star \leftarrow \mathrm{GreedyDPP}(L, k, B, \{\ell_i\})$
$C \leftarrow \mathrm{Serialize}\,(q, \{(ID_i, b_i, e_i) : e_i \in S^\star\}; \mathrm{attribution})$
$\hat{y} \leftarrow G_\phi(C)$
return $(\hat{y}, S^\star)$

5 THEORETICAL ANALYSIS

We analyze the validity of the proposed kernel, its quality–diversity decomposition, and how basis similarity affects the joint selection of passages with similar claims. Proofs are provided in Appendix A.

Lemma 1 (Validity of the BasisDPP kernel). The RBF kernels K^c and K^b in Equations 11–12 are positive semidefinite. Hence, by the Schur product theorem, $K = K^c \odot K^b \succeq 0$. Since W has positive diagonal entries, $L = WKW + \varepsilon I$ is positive definite for $\varepsilon > 0$, and all principal determinants used by the DPP are positive.

Lemma 2 (Quality–diversity decomposition). For $L^{(0)} = WKW$ and any subset S with $\det(K_S) > 0$, the log-determinant decomposes as in Equation 15. Thus, β controls the trade-off between query relevance and basis-aware diversity.

Theorem 1 (Pairwise effect of basis similarity). For two candidates i and j, let $\rho_{ij} = K_{ij}^c K_{ij}^b$. Since $K_{ii} = K_{jj} = 1$, the regularized kernel satisfies

$$\det(L_{\{i,j\}}) = w_i^2 w_j^2 (1 - \rho_{ij}^2) + \varepsilon(w_i^2 + w_j^2) + \varepsilon^2. \tag{23}$$

For fixed quality weights and claim similarity $K_{ij}^c > 0$, the determinant increases as basis similarity K_{ij}^b decreases. For passages with similar claims, high basis similarity increases the redundancy penalty, whereas low basis similarity reduces it. A claim-only kernel assigns the same pair score when quality weights and claim similarity match, regardless of provenance. The basis factor therefore allows selection to distinguish repeated accounts from corroboration when these relations are captured by the learned representations. The experiments evaluate how this mechanism affects evidence selection and downstream verification.

6 EXPERIMENTS

6.1 EXPERIMENTAL SETUP

Datasets. We evaluate BasisDPP-RAG on three evidence-intensive fact-verification benchmarks: AVeriTeC (Schlichtkrull et al., 2023) for heterogeneous Web evidence, NLI4CT (Jullien et al., 2023) for clinical-trial inference, and SciFact (Wadden et al., 2020) for scientific claims and abstracts. These datasets cover complementary evidence settings, and their benchmark-specific basis mappings and candidate-construction protocols are described in Appendix C.

Baselines. We compare zero-shot LLMs, Standard RAG with relevance-based top-k selection, and controlled adaptations of CRAG, Self-CRAG, Self-RAG, AAR, CF-RAG, FIT-RAG, and Speculative-RAG (Yan et al., 2024; Asai et al., 2024; Yu et al., 2023; Qin et al., 2026; Mao et al., 2025; Wang et al., 2025). All methods use the same frozen Llama-2-13B-Chat backbone, candidate pools, context budget, prompt, and decoding configuration; retrieval-based methods differ only in evidence scoring and selection. Zero-shot LLMs use the same generation protocol without retrieval. Appendix B details the adaptations and selector configurations.

Table 1: **Performance comparison on three fact-verification benchmarks.** Values are Smart Exact Match (EM , %). The best baseline score in each column is underlined. Parenthetical arrows report relative improvement over the best baseline in the corresponding column.

Method	AVeriTeC	NLI4CT	SciFact	Avg.
Zero-shot LLMs (no retrieval)				
Llama-2-7B	17.81	15.50	25.68	19.66
Llama-3-8B-Instruct	22.06	25.49	30.17	25.91
Llama-2-13B-Chat	28.13	31.03	34.84	31.33
Standard RAG				
Llama-2-7B	30.14	32.40	34.51	32.35
Llama-3-8B-Instruct	31.53	39.82	41.72	37.69
Llama-2-13B-Chat	40.70	49.68	47.54	45.97
Advanced RAG frameworks				
CRAG	58.37	60.50	63.43	60.77
Self-RAG	64.50	66.37	69.86	66.91
AAR	67.51	70.21	<u>75.32</u>	71.01
CF-RAG	60.03	71.64	73.25	68.31
FIT-RAG	71.87	74.31	74.83	<u>73.67</u>
Self-CRAG	68.10	<u>75.27</u>	73.18	72.18
Speculative-RAG	<u>72.00</u>	73.56	70.94	72.17
BasisDPP-RAG (Llama-3-8B)	**79.52** (↑10.4%)	**82.59** (↑9.7%)	**79.06** (↑5.0%)	**80.39** (↑9.1%)
BasisDPP-RAG (Llama-2-13B)	**82.47** (↑14.5%)	**86.32** (↑14.7%)	**85.73** (↑13.8%)	**84.84** (↑15.2%)

Table 2: **Component ablations of BasisDPP-RAG with Llama-2-13B-Chat.** Values are EM (%). Parentheses show the absolute decrease, in percentage points, relative to the full model.

Dataset	Full BasisDPP-RAG	w/o Evidence Basis	w/o Determinantal Selection	w/o Relevance Quality
AVeriTeC	**82.47**	73.39 (-9.08)	71.64 (-10.83)	76.52 (-5.95)
NLI4CT	**86.32**	70.42 (-15.90)	69.84 (-16.48)	80.26 (-6.06)
SciFact	**85.73**	76.61 (-9.12)	72.20 (-13.53)	79.54 (-6.19)
Average	**84.84**	73.47 (-11.37)	71.23 (-13.61)	78.77 (-6.07)

Implementation. We use frozen Llama-3-8B-Instruct and Llama-2-13B-Chat generators. All retrieval-based methods receive the same six candidates and select at most three passages under a shared context budget, prompt, decoding, and input/output limits. Unless otherwise stated, $\lambda_{\text{set}} = 0.1$, $\lambda_{\text{basis}} = 0.5$, $\tau_{\text{basis}} = 0.1$, and $\beta = 1$, selected on development data. Methods differ only in evidence scoring and selection. We report Smart Exact Match (EM) after canonicalizing predictions; invalid outputs are counted as incorrect. Results are averaged uniformly across the three benchmarks. Evaluation and output protocols are provided in Appendices C.3 and F.

6.2 Main Results

Table 1 reports Smart Exact Match across the three benchmarks. BasisDPP-RAG outperforms Standard RAG and all advanced RAG baselines on every dataset.

Overall Performance. With Llama-2-13B-Chat, BasisDPP-RAG achieves 82.47%, 86.32%, and 85.73% EM on AVeriTeC, NLI4CT, and SciFact, respectively, yielding an average of 84.84%. It improves over FIT-RAG, the strongest baseline on average, by 11.17 percentage points, and exceeds the best baseline on each individual benchmark by more than 10 points. This consistent improvement is compatible with the intended effect of Claim–Basis factorization: suppressing repeated passages from one evidence basis while retaining corroboration from distinct bases.

Across Generator Backbones. BasisDPP-RAG achieves average EM scores of 80.39% with Llama-3-8B-Instruct and 84.84% with Llama-2-13B-Chat, improving over Standard RAG with the corresponding backbone by 42.70 and 38.87 percentage points, respectively. These results suggest that the gains primarily arise from evidence-set selection rather than generator scale alone.

Table 3: **Evidence-selection behavior on NLI4CT with Llama-2-13B-Chat.** EM, IBR, IBP, and basis redundancy are percentages. IBR and IBP measure recall and precision over unique evidence bases. Redundancy is the fraction of selected passages beyond one representative per unique basis, rather than a pair-count statistic; lower redundancy and fewer mean input tokens are preferred.

Variant	EM	IBR	IBP	Redundancy	Input Tokens
Standard RAG	49.68	79.84	67.65	57.76	573.40
Full BasisDPP-RAG	**86.32**	**97.85**	**90.67**	**7.89**	**551.20**
w/o Evidence Basis	70.42	80.15	71.28	38.83	564.10
w/o Determinantal Selection	69.84	88.05	78.39	23.41	561.20
w/o Relevance Quality	80.26	91.21	87.22	18.89	558.60

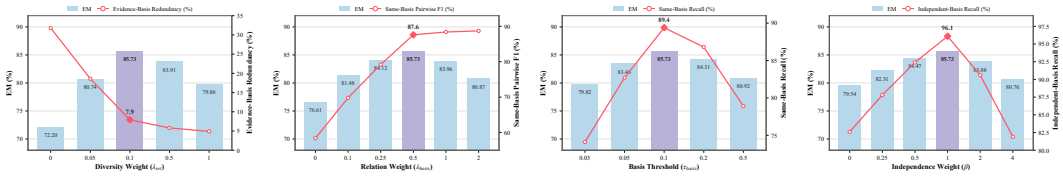


Figure 3: One-at-a-time hyperparameter sensitivity on SciFact with Llama-2-13B-Chat. Bars and the left axis show Exact Match (EM). Lines and the right axis show the corresponding selection diagnostic: basis redundancy for λ_{set}, same-basis F1 for λ_{basis}, same-basis recall for τ_{basis}, and Independent-Basis Recall as the diagnostic for the relevance weight β. The highlighted bars indicate the default settings: $\lambda_{\text{set}} = 0.1$, $\lambda_{\text{basis}} = 0.5$, $\tau_{\text{basis}} = 0.1$, and $\beta = 1$.

6.3 Ablation Analysis

We evaluate the evidence-basis component, determinantal selection, and relevance quality while keeping the generators, candidate pools, context budgets, prompts, and decoding settings fixed. Table 2 reports answer accuracy, and Table 3 reports evidence-selection diagnostics on NLI4CT. Removing determinantal selection, evidence-basis modeling, and relevance quality lowers average EM by 13.61, 11.37, and 6.07 percentage points, respectively, with the largest effects on NLI4CT. Without basis modeling, redundancy increases from 7.89% to 38.83% and IBR/IBP fall from 97.85%/90.67% to 80.15%/71.28%; without DPP selection, redundancy rises to 23.41% and IBR/IBP fall to 88.05%/78.39%. These results show that claim similarity alone cannot distinguish repeated descendants from corroboration across distinct resolved bases, and that basis dependence must be used by a joint, budget-feasible set selector (Kulesza & Taskar, 2012). Removing relevance quality further lowers EM by 6.07 points, confirming that basis-aware diversity must remain coupled with query relevance. Relative to Standard RAG, the full model reduces NLI4CT basis redundancy by 86.3% while using 3.9% fewer mean input tokens, indicating that the improvement comes from a more informative evidence set rather than a larger generator context.

6.4 Hyperparameter Sensitivity Analysis

We perform one-at-a-time sensitivity analyses on SciFact with Llama-2-13B-Chat, varying λ_{set}, λ_{basis}, τ_{basis}, and β while fixing all other settings. Figure 3 compares Smart Exact Match (EM) with the corresponding selection diagnostic.

Set and basis-relation loss weights. Increasing λ_{set} from 0 to 0.1 raises EM from 72.20% to 85.73% and reduces redundancy from 31.8% to 7.9%. Although $\lambda_{\text{set}} = 1$ further reduces redundancy to 4.9%, EM falls to 79.86%. Similarly, increasing λ_{basis} from 0 to 0.5 raises same-basis F1 from 58.4% to 87.6% and EM from 76.61% to 85.73%, whereas $\lambda_{\text{basis}} = 2$ improves F1 only slightly to 88.7% but lowers EM to 80.87%. Thus, improvements in redundancy or pairwise F1 do not necessarily translate into higher answer accuracy.

Basis resolution and relevance quality. Both EM and same-basis recall peak at $\tau_{\text{basis}} = 0.1$, reaching 85.73% and 89.4%, respectively; lower or higher thresholds reduce basis recovery and answer accuracy. Likewise, $\beta = 1$ maximizes both EM at 85.73% and Independent-Basis Recall at 96.1%. These results support the default configuration as a balance between set-level supervision, conservative basis resolution, and query relevance.

6.5 Targeted Failure-Mode Analysis

We evaluate hidden source dependence, context-budget collapse, and echo-induced bias under shared generators, candidate pools, selection budgets, prompts, and decoding settings.

Table 4: **Error rates on targeted RAG failure modes.** Full denotes BasisDPP-RAG, and Reduction is measured relative to Standard RAG. All values are percentages.

Failure Mode	Description	Standard	w/o Basis	w/o DPP	Full	Reduction
Hidden Source Dependence	Paraphrases derived from one source	39.60	28.90	11.50	**10.80**	**72.70**
Context Budget Collapse	Same-source passages displace independent evidence	55.20	31.70	41.80	**12.40**	**77.50**
Echo-Induced Answer Bias	Repeated paraphrases amplify one misleading source	31.80	19.60	22.40	**8.70**	**72.60**

Table 4 reports error rates over the corresponding diagnostic instances, where lower values are better. BasisDPP-RAG achieves the lowest error in all three diagnostics, reducing error by 72.6–77.5% relative to Standard RAG. The reduction is computed as $100(1 - \text{Error}_{\text{BasisDPP-RAG}} / \text{Error}_{\text{Standard RAG}})$, following Equation 46. Removing basis modeling raises hidden-source-dependence error from 10.80% to 28.90%, while removing DPP selection raises context-collapse error from 12.40% to 41.80%; for echo-induced bias, the two ablations increase error to 19.60% and 22.40%, respectively, from 8.70%. These results show that basis modeling identifies shared evidential origins, while DPP selection prevents repeated passages from consuming the context budget.

6.6 Robustness to Adversarial Distraction

We evaluate robustness on NLI4CT by injecting $N_{\text{adv}} \in \{3, 6, 9, 18\}$ semantically plausible but unsupported passages into the shared candidate pool. All methods use the same augmented pool, selection budget, generator, prompt, and decoding configuration. As shown in Figure 4, EM decreases as distractors increase, but BasisDPP-RAG degrades most slowly: from $N_{\text{adv}} = 3$ to 18, Standard RAG, FIT-RAG (Mao et al., 2025), and BasisDPP-RAG lose 36.4%, 24.2%, and 10.1% of their initial EM, respectively. At $N_{\text{adv}} = 18$, BasisDPP-RAG retains 77.43% EM, outperforming FIT-RAG and Standard RAG by 22.64 and 47.81 percentage points, respectively. These results demonstrate greater robustness to the tested distractor injections under a fixed selection budget.

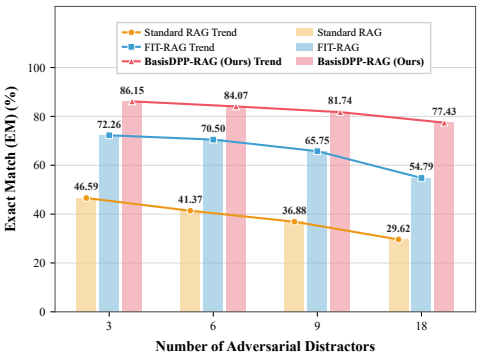


Figure 4: **Adversarial robustness on NLI4CT.**

7 Conclusion and Limitations

RAG should assess not only passage relevance but also whether retrieved evidence provides distinct support. BasisDPP-RAG addresses this issue by representing each passage through its query-conditioned claim and evidence basis, factorizing their similarities in a quality-aware DPP kernel, and selecting a budget-feasible evidence set for generation. The resulting objective suppresses repeated accounts from the same basis while preserving corroboration from distinct bases. Across AVeriteC, NLI4CT, and SciFact, BasisDPP-RAG achieves 84.84% average EM, exceeding the strongest baseline by 11.17 percentage points; on NLI4CT, it reduces basis redundancy from 57.76% to 7.89% and retains 77.43% EM with 18 adversarial distractors. These results show that basis-aware set selection improves both answer accuracy and the evidential quality of retrieved context.

The method assumes that provenance can be recovered and that distinct evidence bases are represented in the candidate pool. Its evaluation is also limited to English fact-verification benchmarks, two frozen Llama backbones, and controlled distractor settings. Future work should address multi-basis provenance, weaker supervision, source reliability, and multilingual or long-form RAG.

Reproducibility Statement

To facilitate reproducibility, we provide an anonymized implementation of the BasisDPP-RAG framework, including the end-to-end selection pipeline, query-conditioned claim extraction, evidence-basis resolution, factorized kernel construction, budgeted DPP inference, training proce-

dures, and generation scripts. The code, configuration files, and scripts required to reproduce the reported experiments are available as supplementary material and at supplementary material. Appendix B describes the model architecture and implementation details, including the training strategy and hyperparameter configurations. Complete proofs of the theoretical results are provided in Appendix A. Appendix C documents the datasets, evidence normalization, candidate construction, and evaluation protocol. The prompt templates, evidence serialization format, output parsing rules, and generation controls are specified in Appendix F. Additional qualitative analyses and evidence-basis resolution details are provided in Appendices D and E.

AI USE STATEMENT

In this work, we used generative AI tools to assist with software implementation and debugging, LaTeX editing, figure and table formatting, and language polishing. Generative AI tools were not used to generate datasets, formulate mathematical claims or proofs, design the experimental methodology, or interpret the experimental results. All AI-assisted outputs were carefully reviewed and verified by the authors, and the authors take full responsibility for the final content of this paper, including all claims, analyses, software, and figures.

REFERENCES

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

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, et al. Improving language models by retrieving from trillions of tokens. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 2206–2240, 2022. URL https://proceedings.mlr.press/v162/borgeaud22a.html.

Jaime Carbonell and Jade Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 335–336, 1998. doi: 10.1145/290941.291025. URL https://doi.org/10.1145/290941.291025.

Xin Luna Dong, Laure Berti-Équille, and Divesh Srivastava. Integrating conflicting data: The role of source dependence. *Proceedings of the VLDB Endowment*, 2(1):550–561, 2009. doi: 10.14778/1687627.1687690. URL https://www.vldb.org/pvldb/vol2/vldb09-pvldb47.pdf.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. Retrieval augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3929–3938, 2020. URL https://proceedings.mlr.press/v119/guu20a.html.

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

Mael Jullien, Marco Valentino, Hannah Frost, Paul O'Regan, Donal Landers, and Andre Freitas. NLI4CT: Multi-evidence natural language inference for clinical trial reports. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 16745–16764, 2023. doi: 10.18653/v1/2023.emnlp-main.1041.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 6769–6781, 2020. doi: 10.18653/v1/2020.emnlp-main.550. URL https://aclanthology.org/2020.emnlp-main.550/.

Alex Kulesza and Ben Taskar. Determinantal point processes for machine learning. *Foundations and Trends in Machine Learning*, 5(2–3):123–286, 2012. doi: 10.1561/2200000044.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen-tau Yih, Tim Rocktaschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638. URL https://aclanthology.org/2024.tacl-1.9/.

Yuren Mao, Xuemei Dong, Wenyi Xu, Yunjun Gao, Bin Wei, and Ying Zhang. FIT-RAG: Black-box RAG with factual information and token reduction. *ACM Transactions on Information Systems*, 43(2):1–27, 2025. doi: 10.1145/3676957.

Huaiyu Qin, Chunyu Wei, Yueguo Chen, and Yunhai Wang. Counterfactual reasoning for retrieval-augmented generation. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/1c078897dc08d46091d0d361d9955c6b-Abstract-Conference.html.

Michael Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. AVeriTeC: A dataset for real-world claim verification with evidence from the web. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-2842. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/cd86a30526cd1aff61d6f89f107634e4-Abstract-Datasets_and_Benchmarks.html.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 31210–31227, 2023. URL https://proceedings.mlr.press/v202/shi23a.html.

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

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 7534–7550, 2020. doi: 10.18653/v1/2020.emnlp-main.609. URL https://aclanthology.org/2020.emnlp-main.609/.

Zilong Wang, Zifeng Wang, Long Le, Huaixiu Steven Zheng, Swaroop Mishra, Vincent Perot, Yuwei Zhang, Anush Mattapalli, Ankur Taly, Jingbo Shang, Chen-Yu Lee, and Tomas Pfister. Speculative RAG: Enhancing retrieval augmented generation through drafting. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=xgQfWbV6Ey.

Fangyuan Xu, Weijia Shi, and Eunsol Choi. RECOMP: Improving retrieval-augmented LMs with compression and selective augmentation. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=mlJLVigNHp.

Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. Corrective retrieval augmented generation. *arXiv preprint arXiv:2401.15884*, 2024. URL https://arxiv.org/abs/2401.15884.

Xiaoxin Yin, Jiawei Han, and Philip S. Yu. Truth discovery with multiple conflicting information providers on the web. *IEEE Transactions on Knowledge and Data Engineering*, 20(6):796–808, 2008. doi: 10.1109/TKDE.2007.190745. URL https://doi.org/10.1109/TKDE.2007.190745.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. Making retrieval-augmented language models robust to irrelevant context. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=ZS4m74kZpH.

Zichun Yu, Chenyan Xiong, Shi Yu, and Zhiyuan Liu. Augmentation-adapted retriever improves generalization of language models as generic plug-in. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pp. 2421–2436, 2023. doi: 10.18653/v1/2023.acl-long.136. URL https://aclanthology.org/2023.acl-long.136/.

A Theoretical Proofs

This appendix provides complete proofs for the results in Section 5. We use the notation of Section 4: K^c and K^b are the claim and evidence-basis kernels, $K = K^c \odot K^b$ is their factorized dependence kernel, $W = \mathrm{diag}(w_1, \ldots, w_n)$ with $w_i = \exp(\beta r_i/2)$, and $L = WKW + \varepsilon I$. We write $L^{(0)} = WKW$ for the kernel before numerical regularization. Because the two base kernels are normalized RBF kernels, their entries lie in $[0, 1]$ and their diagonal entries equal one.

A.1 Proof of Lemma 1: Validity of the BasisDPP Kernel

Proof. The argument has three steps. First, the Gaussian RBF is a positive-definite kernel. Therefore, for any finite candidate pool, its Gram matrices satisfy $K^c \succeq 0$ and $K^b \succeq 0$. This remains true even when two candidates have identical representations, in which case the corresponding Gram matrix may be singular but cannot have a negative eigenvalue.
Second, the Schur product theorem gives

$$K = K^c \odot K^b \succeq 0. \tag{24}$$

This step is central to the factorized construction: combining claim agreement and basis dependence by element-wise multiplication preserves kernel validity, rather than introducing an unconstrained pairwise penalty.
Third, the relevance weights are strictly positive because they are exponentials. For any $\mathbf{v} \in \mathbb{R}^n$,

$$\mathbf{v}^\top L\mathbf{v} = \mathbf{v}^\top WKW\mathbf{v} + \varepsilon\|\mathbf{v}\|_2^2 \tag{25}$$

$$= (W\mathbf{v})^\top K(W\mathbf{v}) + \varepsilon\|\mathbf{v}\|_2^2. \tag{26}$$

The first term is non-negative by Equation 24; the second is strictly positive for every nonzero $\mathbf{v}$ when $\varepsilon > 0$. Hence $L \succ 0$. Every principal submatrix L_S is also positive definite, so $\det(L_S) > 0$. Thus, for $\varepsilon > 0$, the BasisDPP kernel defines a valid L-ensemble and every conditional k-DPP is well-defined. With $\varepsilon = 0$, the same argument gives $L^{(0)} \succeq 0$ and non-negative principal determinants; the corresponding conditional k-DPP is defined only when $Z_k(L^{(0)}) > 0$ (equivalently, $\mathrm{rank}(L^{(0)}) \geq k$). □

A.2 Proof of Lemma 2: Quality–Independence Decomposition

Proof. Let S be any subset for which $\det(K_S) > 0$. Since W_S is diagonal and invertible,

$$L_S^{(0)} = W_S K_S W_S. \tag{27}$$

Determinant multiplicativity gives

$$\det\left(L_S^{(0)}\right) = \det(W_S)^2 \det(K_S) \tag{28}$$

$$= \left(\prod_{i \in S} w_i^2\right) \det(K_S) \tag{29}$$

$$= \exp\left(\beta \sum_{i \in S} r_i\right) \det(K_S). \tag{30}$$

Taking logarithms yields

$$\log\det\Big(L_S^{(0)}\Big) = \beta\sum_{i\in S} r_i + \log\det(K_S), \tag{31}$$

which is Equation **??**.

The same identity clarifies the marginal decision made during inference. For $j \notin S$, the unregularized marginal gain is

$$\Delta(j \mid S) = \log\det\Big(L_{S\cup\{j\}}^{(0)}\Big) - \log\det\Big(L_S^{(0)}\Big) \tag{32}$$

$$= \beta r_j + \log\frac{\det(K_{S\cup\{j\}})}{\det(K_S)}. \tag{33}$$

The first term is the candidate's item-wise relevance contribution; the second is its additional volume in the Claim–Basis geometry. Holding the learned representations fixed, increasing β amplifies only the first term and therefore shifts selection toward item-wise relevance; evidential independence remains represented by the determinant-volume term.

Finally, the stabilizer does not invalidate this interpretation. Because $\det(K_S) > 0$, $L_S^{(0)}$ is positive definite. If $\lambda_1, \ldots, \lambda_{|S|}$ are its eigenvalues, then

$$\log\det(L_S) = \log\det\Big(L_S^{(0)}\Big) + \log\det\Big(I + \varepsilon(L_S^{(0)})^{-1}\Big) \tag{34}$$

$$= \beta\sum_{i\in S} r_i + \log\det(K_S) + \sum_{m=1}^{|S|}\log\left(1 + \frac{\varepsilon}{\lambda_m}\right). \tag{35}$$

The last term is a positive numerical correction that converges to zero as $\varepsilon \to 0$. The exact quality–independence decomposition is therefore recovered in the limit used to define the theoretical objective. □

A.3 Proof of Theorem 1: Separating Repetition from Corroboration

Proof. For normalized kernels, $K_{ii} = K^c_{ii}K^b_{ii} = 1$. Let $\rho_{ij} = K^c_{ij}K^b_{ij}$. The unregularized principal matrix for two candidates is

$$L^{(0)}_{\{i,j\}} = \begin{bmatrix} w_i^2 & w_iw_j\rho_{ij} \\ w_iw_j\rho_{ij} & w_j^2 \end{bmatrix}. \tag{36}$$

Expanding Equation 36 gives

$$\det\Big(L^{(0)}_{\{i,j\}}\Big) = w_i^2w_j^2(1-\rho_{ij}^2), \tag{37}$$

which proves Equation **??**.

Source-dependent repetition. Suppose the candidates express nearly the same query-conditioned claim and resolve to nearly the same basis. For $\delta_c, \delta_b \in [0,1]$, assume $K^c_{ij} \geq 1-\delta_c$ and $K^b_{ij} \geq 1-\delta_b$. Then $\rho_{ij} \geq (1-\delta_c)(1-\delta_b)$, and therefore

$$\det\Big(L^{(0)}_{\{i,j\}}\Big) \leq w_i^2w_j^2\left[1-(1-\delta_c)^2(1-\delta_b)^2\right]. \tag{38}$$

As both discrepancies vanish, the upper bound converges to zero. Thus adding a paraphrased descendant of the same evidential source contributes vanishing joint volume even if both passages have high relevance weights.

Independent corroboration. Now suppose the pair agrees on the query-relevant claim but arises from independent evidence bases. Let $K^c_{ij} \geq 1-\delta_c$ and $K^b_{ij} \leq \eta_b$ for a small $\eta_b \in [0,1]$. Since $K^c_{ij} \leq 1$, we have $\rho_{ij} \leq \eta_b$, giving

$$\det\Big(L^{(0)}_{\{i,j\}}\Big) \geq w_i^2w_j^2(1-\eta_b^2). \tag{39}$$

In the ideal independent-basis limit, $K^b_{ij} \to 0$, this lower bound approaches the maximal pair volume $w_i^2w_j^2$. Moreover, whenever $\eta_b < (1-\delta_c)(1-\delta_b)$, the quality-normalized lower bound for the independent pair exceeds the corresponding upper bound for repetition. The factorized kernel therefore creates a quantifiable separation between repeated support and cross-source agreement, independently of their common multiplicative quality factor.

Extension to evidence sets. Consider a group of exact descendants with identical claim and basis representations. For any third candidate t, their kernel evaluations are identical. The corresponding rows of K are therefore equal, and the rows of $L^{(0)} = WKW$ are proportional after quality scaling. Any subset containing two such descendants is rank deficient and has zero determinant. Repeating a single source cannot increase unregularized DPP mass merely through frequency. An independently sourced passage, by contrast, contributes a non-collinear direction whenever its basis representation is separated from that cluster.

Effect of numerical regularization. For the regularized pair kernel,

$$\det\big(L_{\{i,j\}}\big) = (w_i^2 + \varepsilon)(w_j^2 + \varepsilon) - w_i^2 w_j^2 \rho_{ij}^2 \tag{40}$$

$$= w_i^2 w_j^2 (1 - \rho_{ij}^2) + \varepsilon(w_i^2 + w_j^2) + \varepsilon^2. \tag{41}$$

For exact repetition, only the $O(\varepsilon)$ correction remains; for a pair with low learned basis similarity and the same fixed quality weights, the leading term approaches $w_i^2 w_j^2$. Hence a small stabilizer prevents numerical singularity while preserving the intended preference for separated representations; this statement concerns learned basis geometry, not a guarantee of causal independence. This completes the proof. □

A.4 Computational Complexity

We analyze the selector after the first-stage retriever has produced n candidates. If candidate e_i contains m_i sentences and $M = \sum_i m_i$, sentence-level claim scoring requires M query–sentence encoder evaluations; relevance, claim, and basis representations add at most three candidate-level evaluations. If a shared encoder pass costs C_{enc} and maps to hidden dimension h, the representation stage costs approximately $O((M + n)C_{\text{enc}} + ndh)$. Metadata-based basis resolution is linear in n once provenance indices are available.

Computing all claim- and basis-space distances and constructing K^c, K^b, and K requires $O(n^2 d)$ time and $O(n^2)$ memory. During training, exact k-DPP normalization can be computed from an eigendecomposition of L in $O(n^3)$ time, followed by $O(nk)$ dynamic programming for the elementary symmetric polynomials that define $Z_k(L)$. This cost is incurred only over the moderate candidate pool, not the retrieval corpus.

At inference, greedy MAP selection performs at most k iterations. With incremental Cholesky updates, evaluating and updating marginal gains costs approximately $O(nk^2)$ time and $O(nk)$ auxiliary memory, in addition to the stored $O(n^2)$ kernel. Feasibility checks enforce both the cardinality limit and token budget without changing this asymptotic bound. Only the selected context of at most B tokens is passed to the generator, so BasisDPP-RAG adds no second generation stage and does not increase the generator's asymptotic context length.

B Implementation and Architectural Details

This appendix expands the implementation of the architecture introduced in Section 4. BasisDPP-RAG is implemented as a selector between a first-stage retriever and a frozen generator. The retriever supplies a high-recall candidate pool; the selector constructs query-conditioned claim and query-invariant basis views, evaluates candidates jointly, and exposes only the selected evidence to the generator. This modular boundary is important: the generator never receives the discarded passages, and improvements cannot be attributed to changing the retrieval corpus or adding a second generation stage.

B.1 End-to-End BasisDPP-RAG Pipeline

For each query, the system executes six deterministic interfaces summarized in Table 5. Neural parameters are used only for relevance and representation learning; provenance normalization, budget checking, and context serialization remain explicit operations. The same selector can therefore be paired with either generator backbone without changing the Claim–Basis geometry.

Offline candidate construction retains at most 16 items after adding gold evidence, controlled same-basis descendants, and hard negatives. A common reranking/truncation stage then supplies $n = 6$ candidates to every reported method, and the selector passes at most $k = 3$ passages to the generator. Selection is performed independently for each query because both claim agreement and the feasible token budget are query specific. Basis descriptors, however, are resolved without conditioning on the query so that source identity remains stable across differently worded claims.

Table 5: **End-to-end BasisDPP-RAG implementation.** Each stage exposes a typed output consumed by the next stage, keeping retrieval, set selection, and generation operationally separate.

Stage	Operation	Output
Retrieval	$\mathrm{Retrieve}(q, \mathcal{D})$	Candidate pool $\mathcal{E}(q) = \{e_i\}_{i=1}^{n}$
Claim view	$\mathrm{TopSpan}(e_i; q)$	Extractive claim c_i and claim vector $\mathbf{z}_i^c$
Basis view	$\mathrm{ResolveBasis}(e_i, d_i)$	Canonical basis descriptor b_i and vector $\mathbf{z}_i^b$
Set geometry	$K^c \odot K^b$ with relevance quality	Positive-definite L-ensemble kernel L
Selection	Budgeted greedy DPP MAP	Evidence subset $S^\star$ with $\lvert S^\star \rvert \le k$
Generation	$G_\phi(C(q, S^\star))$	Verdict from the shared evaluation interface

B.2 Query-Conditioned Claim Extraction

The claim view isolates the part of a passage that is relevant to deciding the current query. Each passage e_i is segmented into sentences $\{s_{i1}, \ldots, s_{im}\}$, and the shared cross-encoder scores the paired input $[q; s_{it}]$ using Equation 6. The highest-scoring sentence is always retained. At most one adjacent sentence is added when the top sentence contains or depends on a reference, negation, numerical value, or qualification whose omission would change its interpretation. This produces a short extractive span rather than a generated summary.

Three design choices prevent the claim view from becoming another retrieval score. First, sentence selection is conditioned on q, whereas the original passage remains available to the relevance head. Second, the extracted text is copied verbatim and remains traceable to the evidence identifier; no new claim is synthesized by the generator. Third, the claim projection has parameters separate from the relevance head. Consequently, r_i estimates whether e_i is useful in isolation, while $\mathbf{z}_i^c$ estimates whether two candidates express the same answer-relevant proposition.

The shared encoder processes three input forms: $[q; e_i]$ for relevance, $[q; c_i]$ for claim agreement, and b_i alone for basis dependence. Their outputs are passed through task-specific linear heads, and the claim and basis vectors are ℓ_2-normalized before kernel construction. Normalization makes the diagonal RBF similarities equal to one and prevents vector norm from silently acting as a second quality signal.

B.3 Evidence-Basis Resolution

The evidence basis is the smallest recoverable information origin supporting a passage, rather than the Web page on which the passage happens to appear. For example, two news articles citing the same clinical trial receive different document identifiers but can share one trial basis. Conversely, two studies that reach the same conclusion retain distinct bases even when their wording is highly similar.

We implement $\mathrm{ResolveBasis}$ as a confidence-ranked provenance cascade. Signals are considered from strongest to weakest:

1. persistent identifiers, including normalized DOI, clinical-trial ID, and report or registry number;
2. an explicit citation or canonical URL that resolves to the referenced primary source; and
3. a normalized title–author–date tuple when no persistent identifier is available.

DOI prefixes and letter case are normalized, URL fragments and tracking fields are removed, and composite bibliographic fields are whitespace- and case-normalized before comparison. Semantic similarity alone is never accepted as evidence of shared provenance; it is handled by the claim view instead.

Each proposed link is assigned a resolver confidence. A basis is shared only when the confidence exceeds τ_{basis} and the available identifiers are mutually compatible. Missing metadata, conflicting persistent identifiers, or a score below the threshold trigger the conservative fallback

$$b_i = \texttt{UNRESOLVED:}\mathrm{ID}_i, \tag{42}$$

which is unique to the candidate document. This policy may miss hidden copying, but it avoids the more damaging error of merging independent sources. Text that is recoverable despite missing or conflicting metadata remains in the pool under this unresolved key, whereas unavailable evidence text or index entries are omitted. It also explains why τ_{basis} governs source resolution rather than generic text similarity.

Table 6: **BasisDPP-RAG implementation and inference settings.** Loss and selection parameters correspond to the final method used in the main experiments.

Component	Setting	Value	Role
Selector	Encoder	`all-MiniLM-L6-v2`	Shared text encoder
Selector	Maximum length	160	Selector input tokens
Selector	Projection dimension	96	Claim and basis vectors
Training	Epochs	5	Best development checkpoint retained
Training	Backbone/head LR	$2 \times 10^{-5}/10^{-4}$	AdamW learning rates
Training	Gradient accumulation	4	Effective update interval
Training	λ_{rel}	1.0	Relevance supervision
Training	λ_{claim}	0.2	Claim-pair supervision
Training	λ_{basis}	0.5	Basis-pair supervision
Training	λ_{set}	0.1	Exact k-DPP set loss
Resolution	τ_{basis}	0.1	Minimum basis-link confidence
Selection	β	1	Relevance strength relative to set geometry
Selection	ε	10^{-5}	Numerical stabilization
Selection	Candidate/selected count	$6/3$	Retrieval and context cardinality
Generation	Input/output limit	$3584/96$	Maximum generated context/tokens
Generation	Decoding	Greedy	Deterministic answer generation

The serialized descriptor retains both a canonical key used by the basis head and a human-readable provenance string used in the generator context. The key supports consistent pair labels, while the readable string allows generated citations to expose the source on which a passage depends.

B.4 Kernel Construction and Budgeted DPP Inference

After all candidates are encoded, pairwise squared distances are converted to the RBF kernels in Equations 11–12. The bandwidths σ_c and σ_b are constrained to be positive and are optimized with the selector. The Hadamard product is computed before quality scaling. Hence a pair is strongly redundant only when both its claim and basis similarities are high; a low value in either view prevents independent or complementary evidence from being collapsed.

Relevance scores are converted to positive weights and applied symmetrically, yielding $L = W(K^c \odot K^b)W + \varepsilon I$. Kernel construction and log-determinant evaluation are performed in floating-point precision even when the generator is executed at reduced precision. We symmetrize the numerical kernel before adding εI, which removes round-off asymmetry without changing the intended Gram matrix.

At inference, the selected set is initialized as empty. At iteration t, a candidate is feasible only if adding it respects both $|S| \leq k$ and the evidence-token budget. Among feasible candidates, the algorithm selects the largest marginal gain

$$\Delta_j(S) = \log\det(L_{S\cup\{j\}}) - \log\det(L_S). \tag{43}$$

The procedure terminates when the cardinality is reached, no candidate fits the remaining budget, or no positive marginal gain remains. Incremental Cholesky updates reuse the factorization of L_S rather than recomputing every determinant from scratch. This is a greedy approximation to the budgeted MAP objective and does not guarantee the global optimum.

B.5 Training Strategy and Hyperparameter Configuration

Training examples are grouped by query so that the complete candidate kernel is available for the set-level objective. Item labels supervise relevance; pair labels separately supervise answer-relevant claim agreement and shared evidence basis. In particular, independently corroborating pairs have $y^c_{ij} = 1$ and $y^b_{ij} = 0$. Pairs with unresolved or conflicting provenance are masked from $\mathcal{L}_{\text{basis}}$ rather than treated as independent negatives.

Optimization proceeds in two stages. We first optimize the relevance, claim, and basis heads, establishing stable semantics for the two kernel spaces. We then activate $\mathcal{L}_{\text{set}}$ and jointly optimize the full objective in Equation 21. Because determinants and the exact k-DPP normalizer are differentiable, gradients pass directly through quality weights, RBF bandwidths, and the shared encoder. No policy gradient, straight-through estimator, or generator fine-tuning is required.

Table 6 lists the settings used by the reported experiments. Development data are used for checkpoint and hyperparameter selection; the selected values are then fixed across all three benchmarks.

Table 7: **Dataset characteristics and operational evidence bases.** Split sizes are train/dev/test. The final column identifies the provenance unit used to determine whether two passages depend on the same source.

Dataset	Evaluation focus	Official split	Decision space	Evidence basis
AVeriTeC	Open-web claim verification with source-linked question–answer evidence	3,068/500/1,000	Supported, Refuted, Conflicting, NEI	Canonical source URL
NLI4CT	Inference over single or paired clinical-trial reports	1,700/200/500	Entailment, Contradiction	ClinicalTrials.gov trial ID
SciFact	Scientific-claim verification with sentence-level rationales	809/300/300	Supports, Refutes, NOINFO	S2ORC paper ID

B.6 Generation and Reproducible Execution

The reference implementation targets Python 3.9 or later, NumPy 1.24 or later, PyTorch 2.1 or later, and Transformers 4.40 or later. Determinantal operations are isolated from generator inference, allowing the selector to run in full precision while the frozen Llama generator uses the precision supported by the available accelerator. Randomized decoding is disabled.

For every selected passage, the serialized context contains its evidence ID, human-readable basis descriptor, and original passage text. The reported experiments use the closed-label JSON interface in Appendix F; the citation-aware long-form interface is optional for inspection or deployment and is not used to compute reported EM. Unselected candidates are not appended after selection, because doing so would reintroduce the same source dependence that the DPP removed.

We use frozen Llama-3-8B-Instruct and Llama-2-13B-Chat generators. Each method is evaluated with the same candidate pool, evidence budget, prompt interface, maximum input length, decoding rule, and evaluation split. Advanced RAG baselines are compared as shared-backbone mechanism adaptations under this common protocol. Consequently, differences in the reported experiments are attributable to evidence processing and selection rather than to a larger generator context or method-specific decoding budget.

C Datasets and Evaluation

Appendix B specifies how BasisDPP-RAG transforms a retrieved pool into a budget-feasible evidence set. This appendix completes that description by specifying what constitutes an evidence item and an evidence basis on each benchmark, how controlled candidate pools are built, and how answers and selected sets are scored. The organization follows the same order as the experimental pipeline: benchmark instances are first normalized into provenance-aware candidates, all methods then receive the same retrieval inputs, and answer quality is evaluated together with basis-level diagnostics.

C.1 Dataset Selection and Characteristics

Our evaluation suite spans open-web fact checking, structured clinical evidence, and scientific literature. This domain variation is intentional: the observable form of source dependence changes across corpora, whereas the BasisDPP-RAG interface remains fixed. Table 7 reports official dataset statistics; these values describe the source benchmarks rather than the number of instances retained by a particular experimental run.

AVeriTeC. AVeriTeC contains 4,568 real-world claims collected from 50 fact-checking organizations and pairs verdicts with question–answer evidence and source URLs (Schlichtkrull et al., 2023). Its temporal split and open-web provenance make it useful for studying repeated reporting: multiple pages can be topically relevant while ultimately providing no additional evidential origin. We preserve its four-way decision space and canonicalize annotated source URLs before assigning basis IDs.

NLI4CT. NLI4CT comprises 2,400 statements grounded in 1,000 breast-cancer clinical trial reports from ClinicalTrials.gov (Jullien et al., 2023). Sixty percent of the statements concern one trial and 40% compare two trials; evidence is localized to eligibility, intervention, results, or adverse-events sections. Several facts from one report can therefore agree without constituting independent corroboration. All facts from one trial share a basis, while the primary and secondary trials in comparison examples retain distinct basis IDs.

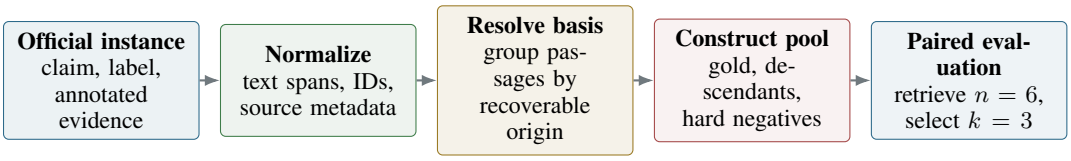


Figure 5: **From benchmark annotations to a shared evaluation pool.** Dataset-specific provenance is normalized before candidate construction; every compared selector then receives the same six-item input pool.

Table 8: **Dataset-specific construction of provenance-aware candidates.** Hard negatives preserve topical or local similarity while lacking gold evidential support.

Dataset	Positive evidence	Hard negatives	Basis assignment
AVeriTeC	Annotated answers concatenated with Boolean explanations and grouped by source	Evidence from other claims with distinct source URLs	Canonicalized URL after normalizing scheme, host, and path and removing query and fragment components
NLI4CT	Evidence sentences reconstructed from annotated indices in the relevant report section	Non-rationale lines from that same section	Clinical-trial ID; paired trials remain separate
SciFact	Annotated rationale sentences grouped by cited abstract	Sentences from abstracts not marked as evidence for the claim	S2ORC paper ID shared by all sentences in one abstract

SciFact. SciFact provides 1,409 expert-written scientific claims paired with 5,183 research abstracts and sentence-level supporting or refuting rationales (Wadden et al., 2020). Claims without a supporting or refuting abstract receive the NOINFO label. Because several rationale sentences may come from one publication, the S2ORC paper identifier—not an individual sentence—defines the evidence basis.

Why this benchmark suite is diagnostic

AVeriTeC tests provenance under heterogeneous Web reporting, NLI4CT tests dependence inside highly structured reports, and SciFact tests repeated rationales within scientific publications. A method that improves all three cannot rely on one corpus-specific metadata convention; it must consistently separate passage relevance from evidential independence.

C.2 Evidence Normalization and Candidate Construction

For each query, preprocessing produces $e_i = (t_i, d_i, b_i, m_i)$, where t_i is the evidence text, d_i is its document identifier, b_i is its canonical evidence-basis identifier, and m_i stores provenance metadata. Figure 5 shows how this representation connects the official benchmark annotations to the common selector interface introduced in Table 5.

The three datasets expose different raw annotations, so the construction rules are specialized only at the normalization boundary. The resulting tuples have the same schema and can be processed by one Claim–Basis encoder. Table 8 summarizes the exact positive, negative, and basis-assignment rules.

Controlled same-basis descendants. Real retrieval pools do not provide enough labeled repetitions to isolate source dependence from lexical relevance. We therefore add up to two controlled surface descendants for each eligible gold candidate during selector training and diagnostic construction. A reporting prefix changes the observed surface context while the evidence core and parent basis ID are preserved. Descendants are marked non-gold and are never counted as new independent support. Offline preparation retains at most 16 items so that gold evidence and controlled descendants are not removed prematurely; the common first-stage interface subsequently reranks and truncates this pool to the six candidates used by every reported method.

Worked basis-assignment example (schematic NLI4CT instance)

Query. Does the intervention reduce severe adverse events relative to the comparator?

Item	Observed passage	Resolved basis	Two-slot decision
e_1	Results section reports fewer severe adverse events	`NCT-A`	Selected
e_2	A summary restates the same trial result	`NCT-A`	Suppressed
e_3	An independent trial reports a compatible reduction	`NCT-B`	Selected

A relevance-only top-k rule can spend both slots on e_1 and e_2. BasisDPP-RAG recognizes that they are descendants of the same trial basis and instead retains e_1 with independently corroborating e_3. The example illustrates the selection mechanism; it is not used as an additional evaluation instance.

Query records, gold evidence, and supervision labels remain in their official source splits. Development data determine checkpoints and $\lambda_{\text{set}}, \lambda_{\text{basis}}, \tau_{\text{basis}}, \beta$; these values are then fixed for the designated evaluation partition. For AVeriTeC only, cross-claim hard negatives are drawn from a shared, unlabeled source-answer pool assembled from the available training and development records; no verdict label is transferred with these negatives. Every method receives the same frozen candidate pool, so this choice affects candidate construction rather than the method-specific comparison.

C.3 Evaluation Protocol and Metrics

Evaluation is paired at the query level. For a given query, every method sees the same six candidates and uses the same generator, evidence-token budget, prompt template, output limit, and greedy decoding rule. Only the evidence-selection function changes. This protocol prevents a stronger retriever, larger context, or more permissive decoding configuration from being mistaken for a selection improvement.

Paired evaluation contract

Retrieval control	Identical candidate IDs and ordering before each selector.
Selection control	The common limit is $k = 3$ under one evidence-token budget.
Generation control	Frozen backbone, prompt, input/output limits, and greedy decoding are shared.
Model selection	Development data choose checkpoints and hyperparameters; the evaluation partition is not used for tuning.

Answer accuracy. The main tables report Smart Exact Match, abbreviated as EM. Unicode NFKC normalization, case folding, punctuation and whitespace normalization are applied before comparison. Predictions are canonicalized within each benchmark's reference label space: AVeriTeC uses Supported/Refuted/Conflicting/NEI, NLI4CT uses Entailment/Contradiction, and SciFact uses Supports/Refutes/NOINFO. The shared JSON interface is mapped to these dataset-specific labels before scoring; a label outside the valid reference space is incorrect. Empty or unparsable outputs are also incorrect. Dataset EM is the mean binary score multiplied by 100, and AVG. is the unweighted macro average over the three benchmarks.

Normalization example

Within the appropriate dataset schema, `Supported`, `SUPPORT`, and `Entailment` map to the canonical SUPPORT decision; `Refuted`, `Contradiction`, and `False` map to REFUTE. Dataset-specific mappings then map SUPPORT, REFUTE, MIXED, and INSUFFICIENT to the valid reference labels; unsupported labels are rejected. This normalization tolerates surface variation without awarding credit to an invalid label.

Table 9: **Evaluation measures and their diagnostic roles.** Answer quality and evidence-set quality are reported separately to avoid treating a correct but source-dependent prediction as fully reliable.

Metric	Preferred direction	Interpretation
Smart EM	Higher	Correctness of the final benchmark-valid canonical decision
Independent-Basis Recall (IBR)	Higher	Coverage of gold evidential origins by the selected set
Independent-Basis Precision (IBP)	Higher	Fraction of selected unique bases that belong to the gold basis set
Selected-basis duplicate fraction	Lower	Fraction of selected passages beyond one representative per distinct basis
Input Tokens	Lower at equal EM	Serialized context cost presented to the frozen generator

Evidence-basis diagnostics. Let S be the selected set, $\mathcal{B}(S) = \{b(e) : e \in S\}$ its unique basis IDs, and $\mathcal{B}^\star$ the gold basis set. We compute

$$\mathrm{IBR}(S) = \frac{|\mathcal{B}(S) \cap \mathcal{B}^\star|}{|\mathcal{B}^\star|}, \qquad \mathrm{IBP}(S) = \frac{|\mathcal{B}(S) \cap \mathcal{B}^\star|}{|\mathcal{B}(S)|}, \tag{44}$$

$$\mathrm{Red}(S) = 1 - \frac{|\mathcal{B}(S)|}{|S|}. \tag{45}$$

Here $\mathrm{Red}(S)$ is the selected-basis duplicate fraction: the fraction of passages beyond one representative per distinct selected basis; it is not a pair-count statistic. It is defined as zero for $|S| < 2$. Queries with no annotated gold basis are excluded from IBR/IBP macro averages; when the selected basis set is empty, IBP is zero if a gold basis exists. All basis metrics are first computed per query and then macro-averaged. Table 9 clarifies why each measure is needed alongside answer EM.

Controlled robustness and failure analysis. For adversarial-distraction experiments, the same semantically plausible but non-supporting passages are injected into every method's pool at $\{3, 6, 9, 18\}$ distractors. They never enter the gold basis set, and all remaining inference settings stay fixed. We report EM at each distraction level. Targeted failure analysis additionally reports the relative error reduction against Standard RAG,

$$\text{Reduction} = 100 \left(1 - \frac{\mathrm{Error}_{\mathrm{BasisDPP-RAG}}}{\mathrm{Error}_{\text{Standard RAG}}}\right). \tag{46}$$

This paired design connects the main results, component analysis, and robustness experiments under one evaluation contract: improvements must be visible not only in final answers, but also in broader independent-basis coverage and lower same-source repetition.

D Qualitative Analysis and Case Studies

The aggregate results in the main paper establish that BasisDPP-RAG improves answer accuracy while reducing same-source repetition. We now inspect how that change arises at the level of one candidate pool and where the mechanism still fails. The analysis follows the interfaces defined in Appendix B and the evaluation contract in Appendix C: evidence text, basis IDs, selected items, and predictions are traced separately. Unless stated otherwise, all numbers below are read from held-out diagnostic logs and do not introduce additional examples into the aggregate results.

D.1 Dissecting Hidden Source Dependence: An NLI4CT Case Study

We select a comparison instance from NLI4CT (Jullien et al., 2023) for which Standard RAG and BasisDPP-RAG receive the same six candidates but produce different evidence sets and verdicts. The example is particularly diagnostic because the statement contains two clauses, each grounded in a different clinical trial. Recovering only one trial is therefore relevant but incomplete.

Table 10: **Candidate groups in the NLI4CT case.** Surface descendants are deliberately repeated observations, not additional evidence bases.

Group	Trial basis	Candidates	Evidence function
Primary	NCT00194779	P_0, P_1, P_2: original + two descendants	Resolves the multi-drug and surgery portion of the statement
Secondary	NCT01743560	S_0, S_1, S_2: original + two descendants	Resolves the two-drug and no-surgery portion of the statement

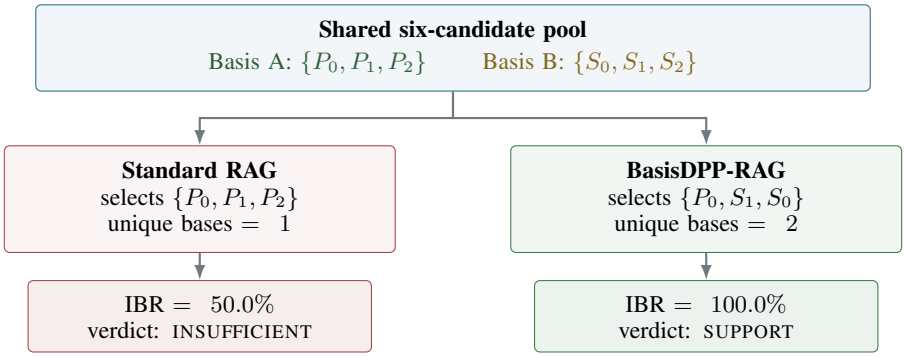


Figure 6: **Selection trace for a source-dependent NLI4CT query.** Relevance ranking spends all three slots on one trial; BasisDPP-RAG covers both trial bases and recovers the gold verdict under the shared semantic interface.

Held-out query and gold decision (shared semantics)

Instance. f8a72a9e... **Section.** Intervention **Gold.** SUPPORT

Claim (paraphrased). The primary trial administers more than seven drugs and includes surgery, whereas the secondary trial administers only two drugs and includes no surgery.

Basis A	NCT00194779: multiple listed agents, including doxorubicin, cyclophosphamide, and paclitaxel, together with definitive breast surgery	Supports the primary-trial clause
Basis B	NCT01743560: everolimus and exemestane are the two listed interventions; no surgery is specified	Supports the secondary-trial clause

For readability, this case uses the shared four-label semantic notation. Here SUPPORT maps to NLI4CT's benchmark label ENTAILMENT; the displayed INSUFFICIENT output has no valid NLI4CT counterpart and is therefore counted as incorrect under the benchmark-specific scoring rule in Appendix C.3.

The controlled pool contains one original evidence item and two surface descendants for each trial. Descendants preserve the evidence core and basis ID but add alternative reporting context, as described in Appendix C.2. Thus all six candidates are relevant at the passage level, yet they represent only two independent origins.

Figure 6 visualizes the resulting selection contrast. Standard top-k exhausts its three slots on Basis A because the three surface forms have closely matched relevance scores. BasisDPP-RAG retains one Basis-A item and allocates the remaining context to Basis B, thereby exposing both clauses of the comparison to the generator.

The exact trace is summarized in Table 11. Both selectors attain perfect Independent-Basis Precision because every selected basis is gold. They differ in recall: Standard RAG covers only one of the two required trials, whereas BasisDPP-RAG covers both. The more informative set also uses fewer prompt tokens because it avoids two long repetitions of the primary-trial evidence.

Table 11: **Per-instance comparison on the held-out NLI4CT case.** Predictions use the shared semantic interface; NLI4CT scoring maps SUPPORT to ENTAILMENT and rejects INSUFFICIENT. IBR, IBP, and redundancy follow the definitions in Equation 44.

Method	Selected basis multiset	IBR	IBP	Red.	Tokens	Prediction
Standard RAG	$\{A, A, A\}$	50.0	100.0	66.7	652	INSUFFICIENT
BasisDPP-RAG	$\{A, B, B\}$	100.0	100.0	33.3	519	SUPPORT

Table 12: **Qualitative pair geometry in the case study.** Relations describe the logged candidate construction; no unobserved numerical kernel scores are imputed.

Candidate pair	Claim relation	Basis relation	Set-level consequence
(P_0, P_1)	Near-identical support	Same trial	Repeated direction; small determinant-volume gain
(P_0, S_0)	Complementary query clauses	Independent trials	Adds a missing evidential direction and expands coverage
(S_0, S_1)	Near-identical support	Same trial	Useful only after the secondary basis is represented
Relevant item vs. off-claim item	Low agreement	Potentially independent	Relevance quality reduces selection of an off-claim item chosen only for diversity

Mechanistic interpretation

The Standard RAG error is not caused by retrieving irrelevant evidence. Every selected passage accurately describes the primary trial. The failure comes from mistaking three views of one basis for complete support of a two-basis comparison. BasisDPP-RAG changes the evidence-set geometry before generation: it preserves relevance while ensuring that the missing secondary-trial premise is represented.

D.2 How Claim–Basis Geometry Changes the Selected Set

The case also clarifies the roles of the two factorized kernels. Claim similarity alone identifies P_0, P_1, P_2 as near-duplicates, but cannot determine whether agreement reflects copied reporting or independent replication. Basis similarity supplies this missing distinction. The Hadamard kernel makes a pair strongly redundant only when the candidates are close in both views, while relevance quality prevents an unrelated but independent passage from being selected merely for diversity.

Why the final set is {A,B,B}, not a hard one-per-basis set

BasisDPP-RAG does not impose a categorical one-per-basis quota. Once Bases A and B are represented, a second item from B may still be selected if its logged marginal gain is positive and it fits the token budget. Thus the observed $\{A, B, B\}$ trace is an outcome of the soft trade-off, not a guarantee; the factorized kernel and relevance scale β jointly determine whether within-basis detail is retained.

D.3 Boundary Cases and Residual Failures

The successful NLI4CT trace should not be interpreted as evidence that provenance-aware selection resolves every RAG error. Table 13 separates failures of basis resolution, retrieval, set construction, and generation. This decomposition is important because each failure requires a different remedy; increasing the DPP diversity weight cannot recover an evidence basis that was never retrieved or repair an incorrect final verdict.

Observed residual error: good evidence structure, wrong verdict

In an auxiliary AVeriTeC development trace (`averitec:dev:229`), the selector covers both gold bases (IBR $= 100\%$, IBP $= 100\%$) with zero selected-source redundancy, yet the generator predicts REFUTE for a gold INSUFFICIENT instance. This example isolates a downstream generation or evidence-to-label mapping error: evidence independence is necessary for reliable generation, but it is not sufficient when the generator maps incomplete support to an overconfident verdict.

Table 13: **Residual failure modes of BasisDPP-RAG.** The final column identifies the appropriate direction for mitigation rather than attributing every error to the selector.

Failure	Observable condition	Consequence	Mitigation direction
Provenance under-resolution	Copied evidence lacks a recoverable DOI, trial ID, citation, or canonical URL	Repeated descendants appear independent	Uncertainty-aware linkage and weakly supervised provenance recovery
Basis over-merging	One publication contains genuinely distinct experiments or analyses	Complementary evidence is suppressed as repetition	Hierarchical or multi-basis assignment
Retriever omission	No independently supporting source enters the candidate pool	DPP can only diversify among incomplete inputs	Provenance-aware first-stage retrieval
Conflicting independent bases	High-quality sources support incompatible decisions	Diversity preserves conflict but cannot adjudicate it	Contradiction-aware synthesis and calibrated abstention
Generator error	Selected set has high IBR/IBP and low redundancy, but the answer mapping is wrong	Structurally sound evidence yields an incorrect verdict	Answer verification and evidence-to-label calibration

Taken together, the qualitative evidence supports a narrower and more mechanistic conclusion than answer accuracy alone. BasisDPP-RAG succeeds when the candidate pool contains the required independent origins and basis resolution makes their dependence visible. Its remaining errors arise when that provenance is hidden, the retriever omits a necessary basis, or the generator misinterprets a structurally sound selected set. These boundaries motivate the provenance and calibration extensions discussed in the limitations section.

E Deep Dive into Evidence-Basis Resolution

Appendix B.3 introduced evidence-basis resolution as the provenance layer between retrieved passages and the factorized DPP kernel. Appendix D then showed why this layer matters: three relevant passages may still cover only one underlying origin. This section makes the resolver operationally transparent. We distinguish the deterministic, benchmark-native identifiers used in the reported experiments from open-corpus resolution, describe how basis labels supervise the representation space, and clarify which notions of independence the resulting kernel can and cannot establish.

E.1 Basis Resolution: Design and Operational Modes

An evidence basis is the smallest *recoverable* provenance unit that the benchmark exposes consistently. It is not defined by textual similarity and does not claim to recover an unobservable causal history. In the reported experiments, resolution is deliberately conservative: a clinical-trial ID, a canonical source URL, or a scientific-paper ID is used when present; otherwise the pair is not used as supervised evidence of shared or independent provenance. Figure 7 summarizes this path.

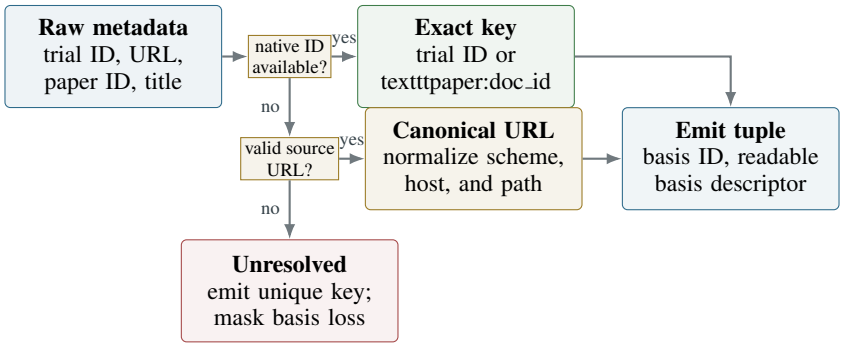


Figure 7: **Deterministic evidence-basis resolution used in the experiments.** Native persistent IDs are preferred; Web evidence is grouped by canonical URL. Recoverable text with unresolved metadata receives a unique key and is masked from basis supervision; unavailable evidence is omitted.

The general resolver interface in Appendix B.3 can accept stronger provenance services upstream, including DOI resolution or citation-graph linkage. Those extensions are useful in open corpora but

Table 14: **Dataset-specific basis resolution in the reported experiments.** The human-readable descriptor is encoded by the basis head; the stable key constructs pair labels and evaluation metrics.

Dataset	Raw field	Stable basis key	Readable descriptor	Missing-field policy
NLI4CT	Primary/secondary trial ID	Exact ClinicalTrials.gov ID	Trial ID plus the relevant report section	Missing metadata receives a unique unresolved key and masks basis supervision; missing evidence text is omitted
AVeriTeC	Annotated source URL	Canonical scheme, host, and path	Canonical source URL	Empty or hostless URLs are marked unresolved and excluded from supervised source groups
SciFact	S2ORC document ID	`paper:doc_id`	Paper ID plus title	Missing metadata receives a unique unresolved key and masks basis supervision; missing document text is omitted

are not required to reproduce the reported benchmark results. The evaluated instantiation uses only fields already present in each dataset, as detailed in Table 14.

Canonicalization example

Observed URL `HTTPS://WWW.Example.org/report/?utm_source=x#results`
Basis key `https://example.org/report`
Operations Lowercase the scheme and host, remove the `www.` prefix, trim the trailing slash, and discard query and fragment components.

The operation merges syntactic variants of one page but does not merge two different URLs merely because their text is similar. Thus canonicalization reduces obvious URL duplication without pretending to solve hidden syndication.

E.2 Basis Supervision and Quality Assurance

The basis head encodes the canonical provenance descriptor/key and never receives the query. This separation prevents query wording from changing source identity. Let $\mathbf{z}_i^b$ be the normalized basis embedding. For known pairs, metadata-derived IDs define $y_{ij}^b = 1$ when $b_i = b_j$ and $y_{ij}^b = 0$ when the known IDs differ. The implemented probability and loss are the same as the basis term in the main objective:

$$p_{ij}^b = \sigma\left(\tau_b(\mathbf{z}_i^b)^\top \mathbf{z}_j^b + \delta_b\right), \qquad \mathcal{L}_{\text{basis}} = \text{BCE}(p_{ij}^b, y_{ij}^b). \tag{47}$$

Pairs containing an `UNRESOLVED` label are masked from this loss. Consequently, missing metadata contributes no false negative asserting that two passages are independent.

E.2.1 Metadata-Derived Pair Labels

Positive pairs arise from repeated observations of one known origin. These include multiple rationale sentences from one SciFact paper, evidence spans from one NLI4CT trial, and controlled descendants that inherit their parent's basis ID. Negative pairs contain two different known IDs, even when their answer-relevant claims agree. The latter case is essential: it teaches the basis head that semantic corroboration across two trials or papers is not source repetition.

E.2.2 Preprocessing Invariants and Masking

Rather than relying on a post-hoc human judgment of source identity, the reported pipeline enforces deterministic invariants during candidate construction. Table 16 lists the checks that connect raw benchmark annotations to the labels used by Equation 47.

Table 15: **Representative basis-pair supervision.** Claim agreement and provenance agreement are labeled independently, preserving the distinction required by the factorized kernel.

Pair	Claim relation	Basis target	Learning effect
Original passage / controlled descendant	Same proposition	Positive	Pull together as repeated observations of one source
Two rationale sentences from one SciFact paper	Complementary or same proposition	Positive	Preserve paper-level dependence across spans
Two NLI4CT trials supporting one comparison	Compatible propositions	Negative	Keep independent clinical origins separated
Two papers reaching the same conclusion	Same answer-level claim	Negative	Preserve genuine cross-paper corroboration
Any pair containing `UNRESOLVED`	Unknown	Masked	Avoid inventing either dependence or independence

Table 16: **Automatic quality controls for basis supervision.** These invariants are applied before selector training and do not use the final answer prediction.

Invariant	Operational check	Failure prevented
Stable key construction	Trial and paper IDs are copied exactly; URLs are canonicalized deterministically	Accidental basis changes across differently worded queries
Descendant inheritance	A controlled descendant clones the parent candidate before changing its evidence ID and surface prefix	Treating repeated reports as independent sources
Missing-metadata mask	`UNRESOLVED` pairs are skipped by the basis BCE	False independence labels from absent provenance
Protected truncation	Gold items and controlled descendants are retained before optional negatives when the temporary pool is capped	Removing the positive dependence pairs needed for training
Query-level determinism	Candidate sampling uses a fixed seed combined with the record ID	Run-to-run changes in pair composition
Label separation	Basis IDs derive from provenance fields, while stance and relevance are stored in separate fields	Leakage from the answer label into source identity

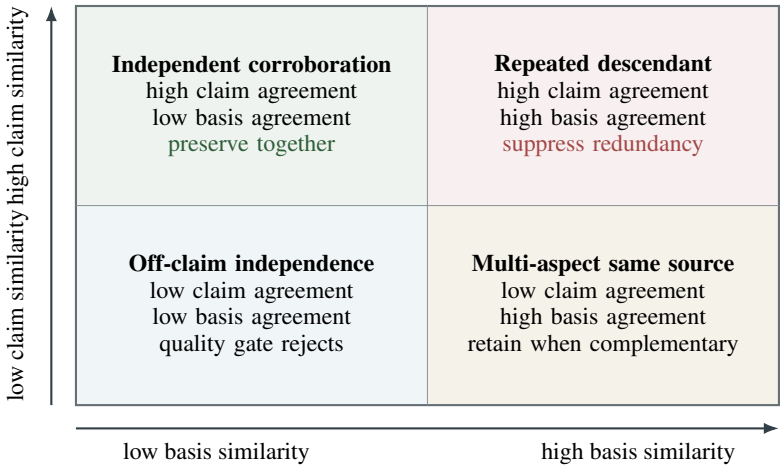


Figure 8: **Claim–Basis pair taxonomy.** The factorized kernel distinguishes independent corroboration from repeated descendants while allowing complementary facts from one source when their claim views differ.

What the supervision asserts—and what it does not

A positive basis label asserts that two passages share the benchmark's recoverable provenance unit. A negative label asserts that their known identifiers differ. Neither label establishes that a source is trustworthy, that two publications are scientifically independent in a causal sense, or that one paper does not copy another without attribution. The basis head learns provenance equivalence under available metadata, not source credibility.

E.3 Learned Claim–Basis Geometry

Basis supervision becomes useful only when combined with the query-conditioned claim head. Figure 8 organizes candidate pairs along these two axes. The upper-left quadrant is the central target of BasisDPP-RAG: independently sourced passages may strongly agree on the answer-relevant claim and should coexist. The upper-right quadrant represents repetition and should contribute little additional determinant volume.

Formally, the redundancy kernel is $K = K^c \odot K^b$. A repeated descendant has large entries in both factors and therefore remains highly similar in K. Independent corroboration can have large claim similarity but small basis similarity, so the product is reduced and the pair can expand the selected-set volume. Conversely, a passage that is independent but irrelevant does not become attractive merely because both similarities are small: its relevance-derived quality weight remains low.

Resolution guarantee and scope

Under correct benchmark identifiers, controlled descendants share a basis label and independently identified trials, URLs, or papers remain separable. This is sufficient to train the geometry tested by the component and redundancy analyses. It is not a guarantee of complete real-world provenance: cross-site syndication, shared upstream datasets, and papers containing multiple independent experiments require hierarchical or multi-basis extensions. The conservative interpretation is therefore *independence with respect to the resolved evidence basis*, not universal causal independence.

This deep dive explains why the basis module improves both selection and diagnostic interpretability. Its contribution is not an additional relevance score; it supplies a stable, query-invariant dependence coordinate. The DPP can then reserve context for independently sourced support without discarding relevant evidence solely because its wording agrees with another passage.

F Prompt Templates and Structured Output Protocol

Section 4.5 defines generation from the selected set, while Appendix B.6 specifies the operational boundary between selection and generation. This section makes that boundary fully reproducible.

Table 17: **Generation interfaces implemented for BasisDPP-RAG.** Selection and provenance are unchanged across interfaces; only the requested answer format differs.

Interface	Serialized fields	Required output	Role
Citation-aware long form	Local evidence tag, extracted claim, original evidence, readable basis, canonical source ID	Verdict plus concise rationale with evidence citations	Inspectable grounded generation and deployment-facing explanations
Closed-label evaluation	Local evidence tag, original passage, canonical source group	One JSON object containing a canonical verdict	Reported Smart-EM experiments under a controlled generation surface

We distinguish two compatible interfaces: a citation-aware long-form interface for grounded responses, and the compact closed-label adapter used in the reported fact-verification experiments. Both expose the same selected passages and resolved source groups; they differ only in the requested output surface.

F.1 Evidence Serialization

Let $S^\star = (i_1, \ldots, i_m)$ be the evidence indices returned by the selector in selection order. The closed-label adapter serializes each selected item as

$$\mathcal{B}_r = [\,\mathrm{E}r;\ \mathrm{Passage} : e_{i_r};\ \mathrm{SourceGroup} : b_{i_r}\,], \qquad r = 1, \ldots, m. \tag{48}$$

The local tag `Er` identifies the displayed block; it is not a new provenance identifier. The source group is the canonical basis key from Appendix E. Consequently, two blocks can have different local evidence tags but still reveal that they descend from one underlying source.

The compact adapter deliberately excludes selector scores, gold indicators, stance labels, and loss targets. The generator sees what was selected and where it came from, but not why the selector preferred it. This prevents relevance or provenance supervision from leaking into the final verdict.

Serialized evidence example

Claim The primary trial used a broader intervention regimen than the secondary trial.

`[E1]` **Passage:** The primary trial reports more than seven drugs and a surgical procedure.
Source group: `NCT00194779`

`[E2]` **Passage:** The secondary trial reports two drugs and no surgery.
Source group: `NCT01743560`

`[E3]` **Passage:** A paraphrased extract repeats the primary-trial regimen.
Source group: `NCT00194779`

Although `E1` and `E3` are separate passages, their identical source group instructs the generator to treat them as one evidential origin. `E2` contributes genuinely independent comparative information.

F.2 Prompt Interfaces

The reported closed-label experiments use one system message and one user message. The wording is shared by all retrieval-based methods; only the selected evidence blocks change. The four-label template below is a shared interface; before scoring, outputs are restricted to the benchmark-specific reference label set listed in Appendix C.3.

Closed-label fact-verification prompt

System

```
You are a rigorous scientific fact-verification model.
Return only one JSON object and no explanation:
{"verdict":"SUPPORT|REFUTE|MIXED|INSUFFICIENT"}.
```

User

```
Determine the verdict using only the supplied evidence.
Treat passages with the same Source group as one underlying
source.
Do not infer that repeated wording constitutes independent
confirmation.
Ignore passages unrelated to the claim; they are not
counterevidence.
Use INSUFFICIENT only when the relevant selected evidence
does not
determine the claim.
Claim: {query}
Selected evidence: {evidence blocks}
```

The provenance instruction is intentionally narrow. It does not tell the generator which source is correct, nor does it expose the DPP score. It only prevents passage multiplicity from being interpreted as independent corroboration. Table 18 gives the decision semantics used by the four-label protocol. A dataset may instantiate only a subset of these labels; unsupported labels are invalid for that benchmark. The citation-aware interface is not used for the reported Smart-EM scores.

Table 18: **Canonical verdict semantics.** The decision depends on the selected evidence set rather than the number of passages expressing one direction.

Verdict	Operational meaning
SUPPORT	The selected evidence directly supports the claim and contains no equally direct unresolved contradiction.
REFUTE	The selected evidence directly contradicts the claim; absence of supporting evidence alone is not refutation.
MIXED	Credible selected evidence supports incompatible directions that cannot be resolved under the available bases.
INSUFFICIENT	The relevant selected evidence does not determine the claim in either direction.

For citation-aware generation, the reference interface adds three constraints: the model may cite only displayed evidence tags, every factual statement in the rationale must cite one of those tags, and an opposite evidential direction must be labeled REFUTE rather than SUPPORT. It requests `Verdict` and `Rationale` fields instead of JSON-only output. The closed-label adapter removes the rationale so that answer verbosity and citation style cannot affect Smart-EM.

F.3 Output Parsing and Label Canonicalization

The generator is instructed to emit valid JSON, but evaluation does not treat minor formatting variation as a semantic error. Figure 9 shows the deterministic parsing cascade. It first reads the `verdict` field from a JSON object. If JSON parsing fails or the field is not recognized, it accepts only an explicit, standalone `Verdict:` field with an exact normalized label. Arbitrary explanation text is never searched for substrings; an output with no recognized label remains invalid.

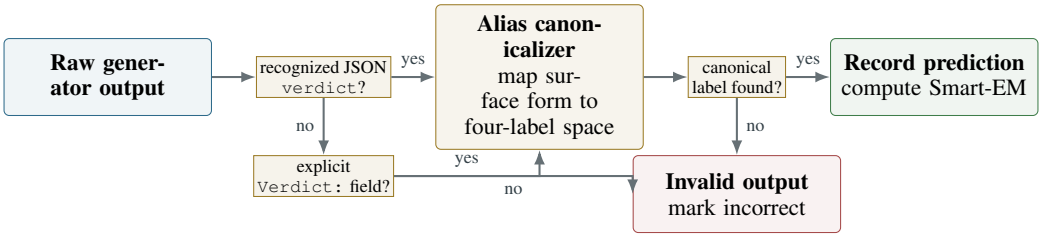


Figure 9: **Deterministic verdict parsing.** Formatting fallbacks recover a recognized label but never infer a decision from evidence or gold annotations.

Table 20: **Controls at the selection–generation boundary.** Values are fixed within each paired comparison.

Control	Fixed information	Bias prevented
Prompt parity	Identical system and user instructions for all retrieval-based methods	Method-specific instructions masquerading as a selector improvement
Context boundary	Only selected passages and their source groups are serialized	Discarded repetitions re-entering the generator context
Supervision exclusion	No gold flag, stance, relevance target, DPP gain, or method name appears in the prompt	Label or method leakage into the final verdict
Decoding parity	Deterministic decoding and a common output cap within each comparison	Sampling variance or unequal generation budgets
Cache identity	Prompt version, model, cardinality, query ID, protocol, and serialized context enter the cache key	Reusing a response for a semantically different evidence set
Failure handling	Unparseable outputs are marked incorrect; no label is imputed from the gold answer	Selective recovery that favors one method

Table 19: **Accepted label aliases in the evaluation implementation.** Alias mapping is applied identically to predictions and reference labels.

Canonical label	Accepted surface forms
SUPPORT	`SUPPORTED`, `ENTAILMENT`, `TRUE`
REFUTE	`REFUTED`, `CONTRADICT`, `CONTRADICTION`, `FALSE`
MIXED	`MIXTURE`, `CONFLICTING EVIDENCE/CHERRYPICKING`
INSUFFICIENT	`NOT ENOUGH EVIDENCE`, `NEI`

Parsing examples

`{"verdict":"REFUTED"}`	JSON field → alias mapping → REFUTE.
`Verdict: not enough evidence`	Explicit-field fallback → INSUFFICIENT.
`The evidence is unclear.`	No recognized alias → invalid output; no verdict is guessed.

The fallback parser is therefore tolerant to formatting noise but conservative about semantics. It cannot convert an explanation into a correct label unless that explanation contains an accepted verdict form.

F.4 Reproducibility and Information-Boundary Controls

Table 20 summarizes the controls that keep final generation comparable across selectors. These controls are especially important for interpreting the component and robustness analyses: an EM change must result from a different selected set, not from a different prompt or decoding budget.

Interpretive boundary

BasisDPP-RAG changes $S^\star$, not the verdict parser. A well-structured evidence set can therefore still produce an incorrect label when the generator misreads direction, sufficiency, or conflict. Conversely, a correct verdict does not prove that the selected passages are independently sourced. This is why the paper reports Smart-EM together with IBR, IBP, and source redundancy, and why Appendix D separates selection failures from residual generation failures.

Taken together, the protocol preserves the main methodological claim at the generation boundary. The selector determines which evidential origins consume context; the shared prompt tells the generator how to interpret those origins; and deterministic canonicalization scores only the resulting

decision. No additional retrieval, self-reflection stage, or hidden source label is introduced after DPP selection.