

Berry-LR: Literature Retrieval for Automatic Scientific Survey Generation via Berrypicking Model

Yajiao Wang^{1,2,6*} Yang Li^{1,2,6*} Xiaotao Wang^{3,4} Xun Feng^{1,2}
Yi Liu^{1,2,6} Mengting Zhang⁵ Xin Lin^{1,2,6} Maodi Hu^{1,2,6} Zhixiong Zhang^{1,2,6†}

¹National Science Library, Chinese Academy of Sciences

²Department of Information Resources Management, School of Economics and Management,
University of Chinese Academy of Sciences ³Institute of Automation, CAS

⁴School of Artificial Intelligence, UCAS ⁵Qiantang Credit Co., Ltd. ⁶Beijing Key Laboratory of
Scientific Artificial Intelligence Research Corpora Data Construction and Governance

Abstract

Automatic scientific survey generation requires a reference set that can support broad topic coverage and citation-worthy reference selection. However, existing systems often approximate this step as one-shot relevance retrieval from a target survey topic, abstract, or generated queries, which may overemphasize papers close to the initial topic wording and miss references needed for broader survey construction. We propose Berry-LR, a literature retrieval framework that models survey reference collection as a berrypicking process. Given a target survey topic, Berry-LR performs move-based search broadening, uses high-confidence retrieved papers as anchors for multi-round retrieval, and records retrieval trajectories as graph evidence for Trajectory-aware Reference Selection. Experiments on SurGE and SurveyBench show that Berry-LR improves reference retrieval, final reference selection, and downstream reference coverage over retrieval and automatic scientific survey generation baselines. Ablations and further analyses suggest that these gains come from move-based search broadening, anchor feedback, and trajectory-aware selection, while many remaining misses are tied to section-local writing contexts.

1 Introduction

The value of a scientific survey lies not in compressing existing papers, but in constructing an interpretable knowledge map of a research area (Torraco, 2005; Snyder, 2019). As the scientific literature continues to grow, large language models have made automatic scientific survey generation increasingly feasible (Wang et al., 2024; Yan et al., 2025; Bao et al., 2025; Su et al., 2025). Yet current systems, despite producing fluent and well-structured long-form text, still fall short of expert-written surveys. A recurring failure mode is that

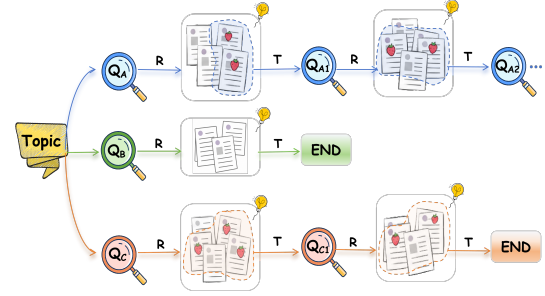


Figure 1: Overview of Berry-LR. Berry-LR broadens retrieval by survey moves, expands through multi-round anchor feedback, and uses retrieval-trajectory evidence to guide trajectory-aware reference selection. R denotes retrieved results and T denotes the intermediate reasoning trace.

the generated survey appears complete, while its underlying literature basis is insufficient, leading to incomplete coverage, weak citations, or unbalanced analysis (Su et al., 2025; Tang et al., 2025).

The reference list of a survey is not peripheral metadata; it forms the evidential backbone of the entire article. It determines which research directions are included, which methods are treated as central, and which neighboring areas provide context and inspiration (Torraco, 2005; Snyder, 2019; Bao et al., 2025). If the reference set misses foundational work, classic methods, important resources, or prior taxonomies, the downstream generator can hardly recover these omissions through language modeling alone. The literature collection stage therefore plays a critical role in shaping the quality of the generated survey. Existing automatic scientific survey generation and deep research systems have recognized the importance of retrieval, often combining it with outline planning, section generation, iterative revision, or citation expansion (Lewis et al., 2020; Shao et al., 2024; Asai et al., 2026; Yan et al., 2025; Chen et al., 2025; Li et al., 2025). However, most systems still rely on one-shot relevance-based retrieval, or embed retrieval

*These authors contributed equally to this work.

†Corresponding authors: zhangzx@mail.las.ac.cn

into local writing procedures such as finding evidence for an outline node, supplementing a section, or expanding context during generation (Yan et al., 2025; Chen et al., 2025; Go et al., 2026; Shi et al., 2025). Such localized retrieval fails to model the global reference set and may narrow the survey’s scope during initial outline generation. Moreover, retrieval signals are often discarded: which query retrieved a paper, whether it appeared across multiple survey moves, and whether it was introduced through an anchor paper. Consequently, literature collection remains limited to topical similarity with the initial query, overlooking lexically distant but highly relevant papers essential for the survey.

Our central observation is that expert survey literature collection resembles berrypicking rather than a single relevance-based retrieval process. Bates’ berrypicking model characterizes search as an incremental, interactive process where users discover cues, reformulate queries, and gather information along multiple paths (Bates, 1989). Similarly, researchers start with a broad topic, identify seed papers, expand along revealed concepts, methods, tasks, or neighboring areas, and monitor topic drift to avoid narrow sub-fields (Snyder, 2019; Agarwal et al., 2025). This process requires broadening the search space to cover the diverse survey knowledge roles needed, while preserving discovery paths so that the retrieval trajectory itself can serve as evidence for citation-worthiness.

Based on this observation, we propose Berry-LR, a literature retrieval framework for automatic scientific survey generation using a berrypicking graph model. As shown in Figure 1, Berry-LR first performs move-based search broadening from the structure of survey writing, decomposing retrieval around the target topic into multiple survey moves. It then performs multi-round anchor-driven retrieval, where high-confidence papers guide subsequent retrieval directions and drift-mitigation mechanisms reduce topical drift. Unlike methods that return only a flat ranked list, Berry-LR records a retrieval evidence graph capturing which queries retrieved each candidate, which survey moves it appeared in, which retrieval rounds it survived, whether it served as an anchor, and whether it overlapped with a baseline retrieval result. These trajectory signals define a trajectory-derived prior for reference selection. The subsequent trajectory-aware selector evaluates each candidate’s marginal contribution relative to the currently selected set, producing a compact, balanced, and functionally

diverse reference set.

Our contributions are as follows:

- We propose Berry-LR, a literature retrieval framework expanding a survey topic into a broader citation space via move-based search broadening and multi-round anchor feedback.
- We model retrieval trajectories as graph-structured provenance evidence and use the resulting prior to guide trajectory-aware reference selection based on each paper’s marginal contribution.
- Experiments show that Berry-LR improves both early reference recall and final reference-set quality over existing retrieval and selection baselines.

2 Related Work

Automatic scientific survey generation systems.

Automatic scientific survey generation builds on earlier work on scientific multi-document summarization and related-work generation, which studies how to select, organize, and synthesize information from cited papers (Hu and Wan, 2014; Lu et al., 2020; Li et al., 2022; Kasanishi et al., 2023). More recent work further emphasizes reference retrieval, structured related work, and taxonomy or hierarchy construction for literature-review support, showing that literature synthesis requires organizing papers by topics, roles, and relations (Shi et al., 2023; Nishimura et al., 2024; Zhu et al., 2023; Hsu et al., 2024). Building on this line, recent LLM-based systems have shifted from section-level related-work generation to end-to-end automatic scientific survey generation. AutoSurvey decomposes survey writing into retrieval, outline generation, subsection drafting, integration, and iterative evaluation (Wang et al., 2024); SurveyX introduces attribute trees to organize retrieved references before generation (Liang et al., 2025); SurveyForge combines outline heuristics with a memory-driven scholar navigation agent (Yan et al., 2025); and Survey-Gen studies quality-aware scientific survey generation with large-scale survey and reference metadata (Bao et al., 2025). Further systems explore step-by-step generation, iterative workflows, personalized interaction, memory-guided writing, multi-agent collaboration, hierarchical citation graphs, reflective refinement, outline-specific modeling, and multi-view structure generation (Lai et al., 2024; Zhang et al., 2025; Wen et al., 2025; Chen et al., 2025; Go et al., 2026; Shi et al., 2025; Nguye et al., 2025; Ma et al., 2025; Liu et al., 2026). These

methods improve outline quality, coherence, and downstream writing, but their retrieval modules are usually treated as an upstream candidate source or as local evidence retrieval during generation. Complementing these approaches, Berry-LR focuses on pre-writing, survey-level reference selection to construct a compact, supportive reference set for the entire article.

Benchmarks and evaluation for survey generation. Several recent resources make automatic scientific survey generation more measurable. SurveyGen provides a large-scale collection of human-written surveys and cited references (Bao et al., 2025); SurGE introduces a large retrieval corpus and multi-dimensional evaluation for scientific survey generation (Su et al., 2025); and SurveyForge introduces SurveyBench and Survey Assessment Metrics for evaluating generated surveys across reference, outline, and content quality (Yan et al., 2025). Related benchmarks further evaluate deep-research agents, discipline-aware survey generation, and retrieve and organize evidence into expert taxonomies (Li et al., 2025; Guo et al., 2026; Zhang et al., 2026). These benchmarks consistently highlight citation coverage, reference quality, and evidence organization as major bottlenecks.

3 Method

3.1 Overview

Berry-LR is designed for the literature collection stage of automatic scientific survey generation. Given a target survey topic x and a paper corpus $\mathcal{D}$, the goal is to construct a reference set $\mathcal{S} \subset \mathcal{D}$ that can support the subsequent survey writing process.

The key assumption behind Berry-LR is that survey references should be selected as a set, rather than as independent top-ranked retrieval results. A robust reference set should cover key writing dimensions, incorporate papers from diverse search paths, and avoid narrow topical clustering. Berry-LR starts from a target survey topic, expands the search space through survey moves, follows high-confidence papers as anchors, and saves the trajectory for subsequent reference selection.

Formally, Berry-LR outputs a set of K papers:

$$\mathcal{S} = \{d_1, \dots, d_K\}, \quad d_i \in \mathcal{D}.$$

The method consists of three operational components: move-based search broadening, anchor-based berrypicking retrieval, and trajectory-aware

reference selection. Retrieval trajectory evidence connects the retrieval process with the final selection stage. Figure 2 provides an overview of the running example, the main pipeline, and the trajectory visualization.

3.2 Move-based Search Broadening

A single query derived from the target survey topic often retrieves papers that are topically close to the input, but may miss papers needed for different parts of a survey. To broaden the search, Berry-LR decomposes the retrieval intent according to common survey-writing moves:

$$\mathcal{M} = \{\text{background, purpose, methods, results, limitations, conclusion}\}.$$

These moves follow work on genre and scientific discourse analysis, where scientific texts are often modeled through rhetorical roles such as background, goals, methods, results, and implications (Teufel, 1999; Liakata, 2010). Here, they serve as general retrieval branches rather than discipline-specific categories.

For each move $m \in \mathcal{M}$, Berry-LR generates a small set of move-conditioned queries:

$$\mathcal{Q}_m^{(1)} = \{q_{m,1}^{(1)}, \dots, q_{m,n}^{(1)}\}.$$

Each query is conditioned on both the target survey topic and the corresponding move, giving Berry-LR multiple structured entry points into the corpus. The detailed prompt templates are provided in Appendix B.

Given a query q , the retriever returns a ranked list $R(q)$. For each move, the results from its queries are merged by reciprocal rank fusion (RRF) (Cormack et al., 2009):

$$s_m^{(r)}(d) = \sum_{q \in \mathcal{Q}_m^{(r)}} \frac{\mathbb{I}[d \in R(q)]}{k_{\text{RRF}} + \text{rank}_q(d)}.$$

Here, $\text{rank}_q(d)$ is the rank of paper d in the result list of query q , and k_{RRF} is a smoothing constant. The fused score $s_m^{(r)}(d)$ measures how strongly d is supported by the queries of move m in retrieval round r .

3.3 Anchor-based Berrypicking Retrieval

Move-based broadening provides the initial search space, but survey writers rarely stop after a fixed set of initial queries. They inspect promising papers, extract new terms or directions from them, and

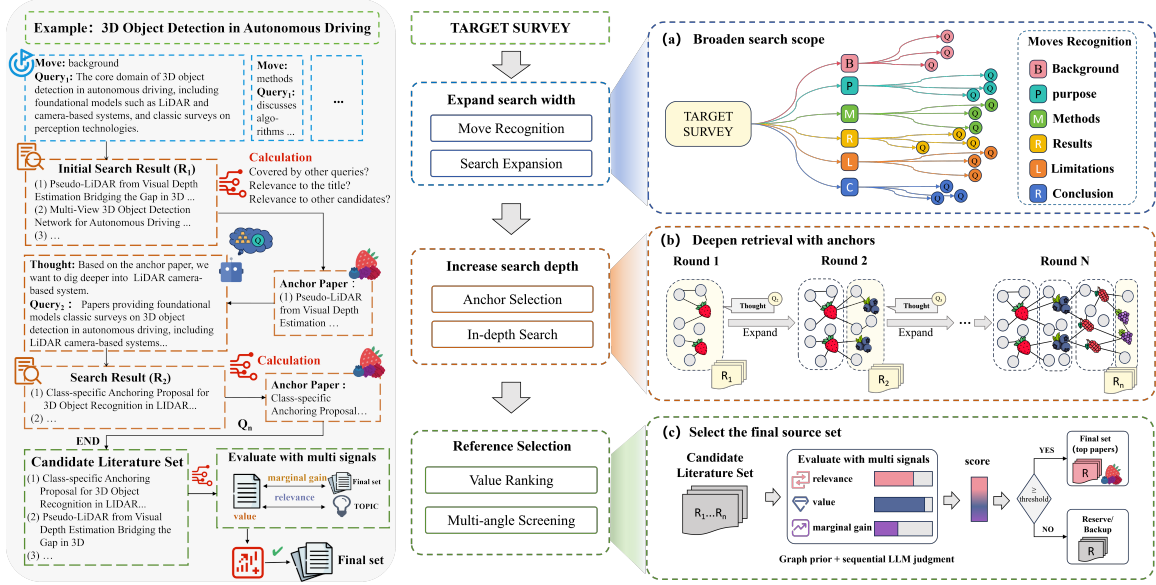


Figure 2: Method overview of Berry-LR. The left panel gives a running example; the middle panel summarizes the pipeline; the right panel visualizes move-based broadening, anchor feedback, and trajectory-aware reference selection.

continue searching. Berry-LR approximates this process with anchor-based berrypicking.

At round r , for each move m , Berry-LR first obtains a fused candidate list $\mathcal{C}_m^{(r)}$ from the move-conditioned retrieval results. A good anchor should be relevant to the move, highly ranked, and supported by multiple queries under the same move. We compute an anchor score:

$$a_m^{(r)}(d) = \lambda_1 \text{sim}_m(d) + \lambda_2 \tilde{s}_m^{(r)}(d) + \lambda_3 \text{cov}_m^{(r)}(d) + \lambda_4 \text{titleSim}_x(d).$$

The four terms measure move relevance, normalized fused retrieval score, query coverage under the move, and lexical similarity to the target survey topic. Only papers whose anchor score exceeds a threshold enter the eligible anchor pool:

$$\mathcal{A}_m^{(r)} = \{d \in \mathcal{C}_m^{(r)} : a_m^{(r)}(d) \geq \tau_{\text{anchor}}\}.$$

Berry-LR then selects the top- L papers from this eligible pool as temporary anchors:

$$\tilde{\mathcal{H}}_m^{(r)} = \text{TopL}_{d \in \mathcal{A}_m^{(r)}} a_m^{(r)}(d).$$

Starting from the second retrieval round, drift control is applied to these temporary anchors. If the eligible pool is empty or drift control triggers, we set $\mathcal{H}_m^{(r)} = \emptyset$ and stop expanding move m . Otherwise, $\mathcal{H}_m^{(r)} = \tilde{\mathcal{H}}_m^{(r)}$. Additional implementation details are given in Appendix A.2.

If $\mathcal{H}_m^{(r)}$ is empty, Berry-LR stops expanding move m . Otherwise, it generates the next-round queries using the target survey topic, the current move, previous queries, and the selected anchors:

$$\mathcal{Q}_m^{(r+1)} = g_\theta(x, m, \mathcal{Q}_m^{(\leq r)}, \mathcal{H}_m^{(r)}).$$

Here, g_θ denotes the query generation procedure. This allows retrieval to move beyond the wording of the initial topic while remaining tied to the survey move that motivated the search.

Because anchor feedback can reinforce early bias, the thresholded anchor pool and drift-control rule keep the search exploratory but bounded. The drift-control rule and prompt design are described in Appendix A.2 and Appendix B.

3.4 Retrieval Trajectory as Evidence

Berry-LR preserves the retrieval process rather than collapsing it into a flat candidate list. For each candidate paper d , we construct an evidence vector:

$$\mathbf{e}(d) = [\mathbf{e}_{\text{ret}}(d); \mathbf{e}_{\text{traj}}(d); \mathbf{e}_{\text{anc}}(d); \mathbf{e}_{\text{agree}}(d)].$$

The four groups capture retrieval strength, trajectory diversity, anchor participation, and agreement between berrypicking retrieval and direct topic-based retrieval. A paper reached across multiple moves, rounds, and query paths gets stronger evidence than one retrieved only by a single local query.

After normalization, Berry-LR maps the evidence vector into a citation-worthiness score:

$$c(d) = \mathbf{w}^\top \phi(d),$$

where $\phi(d)$ denotes the normalized evidence features and $\mathbf{w}$ is the feature weight vector. Sorting candidates by $c(d)$ yields a trajectory-derived prior ranking:

$$\mathcal{C} = [\hat{d}_1, \hat{d}_2, \dots, \hat{d}_N].$$

Detailed feature definitions are provided in Appendix A.3.

3.5 Trajectory-aware Reference Selection

The trajectory-derived prior ranking identifies papers with strong retrieval evidence, but the final reference set should also avoid redundancy, a concern central to diversity-aware reranking and summarization (Carbonell and Goldstein, 1998). A paper that is useful in isolation may contribute little if the selected set already contains several similar papers, while a lower-ranked paper may be useful if it fills a missing part of the survey.

Berry-LR therefore selects references sequentially. At step i , the model considers candidate paper $\hat{d}_i$ together with the current selected set $\mathcal{S}_{i-1}$. It estimates three signals:

$$\mathbf{z}_i = [z_{\text{rel}}(\hat{d}_i), z_{\text{val}}(\hat{d}_i), z_{\text{gain}}(\hat{d}_i | \mathcal{S}_{i-1})].$$

Here, z_{rel} measures topical relevance, z_{val} measures the intrinsic value of citing the paper, and z_{gain} measures its marginal contribution with respect to the already selected papers. These signals are obtained as structured numeric LLM judgments and normalized to a common scale. The corresponding prompt template is provided in Appendix B.

We combine the set-conditioned score with the trajectory-derived prior:

$$\pi_i = \alpha s_{\text{LLM}}(\hat{d}_i | \mathcal{S}_{i-1}) + (1 - \alpha) s_{\text{graph}}(\hat{d}_i).$$

Here, s_{LLM} is derived from the three decision signals, while s_{graph} is derived from the citation-worthiness score and the trajectory-derived prior rank. Both terms are normalized selection scores rather than calibrated probabilities.

Berry-LR accepts $\hat{d}_i$ if the current set has fewer than K papers and $\pi_i \geq \gamma$; otherwise, the candidate is skipped. If fewer than K papers are selected after one pass, Berry-LR fills the remaining slots with the unselected candidates that have the highest final selection scores. The final reference set $\mathcal{S}$ is

therefore guided by both the retrieval trajectory and the marginal value of each paper in the context of the selected reference set.

4 Experiments

We evaluate Berry-LR on reference retrieval and selection for automatic scientific survey generation. Experiments test whether Berry-LR recovers references more effectively than topic-centered retrieval and whether trajectory-derived evidence improves final reference selection.

4.1 Experimental Setup

Datasets. We evaluate Berry-LR on SurGE and SurveyBench. SurGE contains 205 verified surveys and 1,086,992 candidate papers, with each survey’s reference list used as the gold set for retrieval evaluation (Su et al., 2025). SurveyForge’s SurveyBench provides a complementary setting for evaluating whether selected references can support downstream survey generation (Yan et al., 2025).

SurGE retrieval and selection. On SurGE, we report Recall@k under different candidate budgets. Berry-LR is evaluated using two retrieval backends: BM25 and the RoBERTa retriever from SurGE (Su et al., 2025). Unless otherwise specified, Berry-LR refers to the BM25-based version. Since Berry-LR starts from the survey title and expands retrieval internally, we compare baselines under both title-only and title+abstract settings.

We further evaluate final reference selection using precision, recall, and macro-F1 against the gold reference set. Retrieval baselines directly select the top 50 papers from their own ranked lists. Generation baselines and Berry-LR use the controlled Berry-LR candidate pool: the generation baselines cite from this pool, whereas Berry-LR applies trajectory-aware reference selection to it.

SurveyBench evaluation. For SurveyBench, we adopt the SurveyForge protocol to report Input and Reference Coverage (Yan et al., 2025). Input Coverage measures the overlap between the input literature pool and benchmark references, while Reference Coverage measures the overlap between cited references and benchmark references. We set a 700-paper input budget and follow SurveyForge’s per-topic reference budget for final references.

Baselines. For SurGE retrieval, we compare with BM25 (Robertson and Zaragoza, 2009), BGE (Xiao et al., 2024), SPECTER2 (Singh et al., 2023),

Query Input	Method	R@20	R@50	R@100	R@150	R@200	R@500	R@1000	R@1500
Title	BM25	0.0363	0.0645	0.0971	0.1255	0.1406	0.2015	0.2629	0.2992
	BGE-large-en	0.0268	0.0507	0.0780	0.0975	0.1161	0.1907	0.2560	0.3041
	SPECTER2	0.0447	0.0769	0.1176	0.1448	0.1712	0.2503	0.3152	0.3631
	OpenScholar retriever	0.0491	0.0805	0.1190	0.1427	0.1587	0.2313	0.3098	0.3624
	RoBERTa retriever	0.0350	0.0560	0.0881	0.1119	0.1306	0.2070	0.2825	0.3367
Title+ Abstract	BM25	0.0315	0.0643	0.0855	0.1011	0.1122	0.1613	0.2083	0.2397
	BGE-large-en	0.0592	0.0997	0.1459	0.1880	0.2143	0.3011	0.3752	0.4234
	SPECTER2	0.0536	0.0997	0.1495	0.1814	0.2080	0.2976	0.3756	0.4243
	OpenScholar retriever	0.0496	0.0870	0.1155	0.1397	0.1587	0.2309	0.3076	0.3587
	RoBERTa retriever	0.0597	0.1050	0.1586	0.1926	0.2209	0.3292	0.4361	<u>0.5035</u>
Title	Berry-LR (BM25)	0.0823	<u>0.1363</u>	<u>0.1936</u>	<u>0.3001</u>	<u>0.3341</u>	<u>0.4052</u>	<u>0.4453</u>	0.5025
	Berry-LR (RoBERTa retriever)	<u>0.0774</u>	0.1710	0.2731	0.3471	0.3854	0.4393	0.5058	0.6095

Table 1: Gold reference recall of retrieval models. Recall@k is computed over the top- k retrieved papers. Best and second-best results are bolded and underlined.

OpenScholar-based retrieval (Asai et al., 2026), and the RoBERTa retriever from SurGE (Su et al., 2025). For reference selection and SurveyBench evaluation, we compare with retrieval-based selectors and automatic survey generation systems, including AutoSurvey (Wang et al., 2024), SurveyForge (Yan et al., 2025), and SurveyGen (Bao et al., 2025). Additional details on metrics, budgets, and implementation are provided in Appendix A.1.

4.2 Main Experiments

SurGE reference retrieval. Table 1 reports reference retrieval results on SurGE. Overall, Berry-LR consistently improves reference recall by changing the retrieval process, rather than simply enriching the initial query. With BM25, it uses only the survey title yet outperforms all title-based baselines across retrieval depths, achieving 0.3001 recall at R@150, more than doubling the strongest baseline.

Comparisons with title+abstract baselines show a similar trend. Although Berry-LR (BM25) does not use abstract inputs, it outperforms the strongest title+abstract baseline at most retrieval depths and remains comparable at R@1500. In particular, at R@150 it reaches 0.3001 recall, compared with 0.1926 from the strongest title+abstract baseline. This suggests that the gains come from move-based broadening and anchor-based berrypicking, which help find hard-to-retrieve references, rather than merely from additional input.

Berry-LR also benefits from a stronger retrieval backend. When instantiated as Berry-LR (RoBERTa retriever), it achieves the best recall at almost all retrieval depths and reaches 0.6095 at R@1500. This indicates that Berry-LR is not tied to a specific retriever. Even a lexical retriever ben-

efits from the expanded retrieval trajectory, and a supervised scientific retriever further improves the quality of the candidate set.

SurGE reference selection. Table 2 reports final reference selection results on SurGE. Under this budget-controlled protocol, Berry-LR achieves top precision (0.1849), recall (0.1721), and macro-F1 (0.1539) against both retrieval-based selectors and automatic survey generation methods.

Berry-LR substantially outperforms all retrieval-based selectors, including the strongest RoBERTa dense retriever (0.1119 precision, 0.1050 recall, 0.0881 macro-F1), proving that relevance ranking alone cannot yield the optimal reference set.

It similarly surpasses survey generation baselines, beating SurveyGen’s top precision (0.1644) and macro-F1 (0.1305), as well as SurveyForge’s top recall (0.1362). Retrieval baselines use their own top-50 ranked outputs, while generation and selection methods use the controlled Berry-LR candidate pool. The results suggest that Berry-LR’s trajectory-aware selector better identifies compact, gold-aligned reference sets.

SurveyBench results. Table 3 reports the reference-quality results on SurveyBench. The results reveal a clear distinction between retrieval-stage input coverage and final reference coverage. SurveyForge achieves the highest Input Coverage under both generator settings, indicating its 700-paper input pool has stronger raw overlap with the benchmark references. In contrast, Berry-LR achieves the highest Reference Coverage after trajectory-aware reference selection and survey generation.

This affirms Berry-LR’s design: rather than

Model	Precision	Recall	Macro-F1
<i>Retrieval methods</i>			
BM25	0.0381	0.0643	0.0363
SPECTER2	0.0779	0.0997	0.0694
OpenScholar	0.0593	0.0870	0.0558
BGE-large-en	0.0845	0.0997	0.0739
RoBERTa dense retriever	0.1119	0.1050	0.0881
<i>Automatic scientific survey generation methods</i>			
AutoSurvey	0.0940	0.0658	0.0606
SurveyForge	0.1118	0.1362	0.1162
SurveyGen	0.1644	0.1082	0.1305
Berry-LR	0.1849	0.1721	0.1539

Table 2: Reference selection results on SurGE. Retrieval methods use the top-50 papers from their own ranked lists. Survey generation methods and Berry-LR selection use the controlled Berry-LR candidate pool; generation methods use Qwen-3.5-Plus to cite references from that pool.

Method	Model	Reference Quality	
		Input Cov.	Reference Cov.
Human-Written	-	-	0.6294
AutoSurvey	Claude-3 Haiku	0.1153	0.2341
SurveyForge		0.2231	0.3960
Berry-LR		0.1432	0.5305
AutoSurvey	GPT-4o mini	0.0665	0.2035
SurveyForge		0.2018	0.4236
Berry-LR		0.1397	0.4937

Table 3: SurveyBench reference quality comparison across different methods and models. The best result within each model group is shown in bold.

maximizing raw overlap, it selects references highly likely to provide citation-level evidence. For example, under Claude-3-Haiku (Anthropic, 2024), Berry-LR improves Reference Coverage from 0.3960 to 0.5305 over SurveyForge; under GPT-4o-mini (OpenAI, 2024), it improves from 0.4236 to 0.4937. These results suggest that survey-oriented literature retrieval should not only retrieve a broad input pool, but also identify a compact and citation-worthy source set.

The gap between Input Coverage and Reference Coverage indicates that retrieval-stage overlap alone does not guarantee citation-level coverage. Effective reference selection benefits from evidence about how papers are discovered, whether they are supported by multiple retrieval paths, and how much marginal value they add to the selected set.

4.3 Ablation Study

Table 4 reports ablation results on SurGE. All ablations are conducted with Berry-LR based on BM25, so the comparison focuses on the proposed berrypicking components rather than on the underlying retriever. We examine move-based search broadening, multi-round anchor feedback, and retrieval-trajectory evidence to test whether the gains come from structured berrypicking rather than issuing more queries or running more retrieval rounds.

Move-based search broadening. We first show survey moves are essential for broadening retrieval: removing them drops R@1500 from 0.5025 to 0.3254, proving topic-centered signals alone are insufficient. A generic append-retrieve variant generating queries from random perspectives reaches 0.4173 R@1500, better than topic-centered retrieval but still below the full model. This proves Berry-LR’s gains stem not merely from a larger query budget, but from a move structure that expands retrieval while maintaining scope alignment.

Anchor feedback. We next study whether multi-round anchor feedback contributes beyond the initial move-conditioned retrieval. The one-round variant obtains 0.4099 R@1500, indicating that first-round move-conditioned queries already provide a strong candidate set. The full model improves to 0.5025 R@1500, suggesting that later rounds add useful references when retrieved papers are used as anchors. However, anchor feedback must be constrained: Random anchors and Top-ranked anchors, under the same maximum of five anchors per move-round, reach only 0.4329 and 0.4426 R@1500, respectively. Random anchors can introduce unstable or overly narrow feedback, while rank-only anchors tend to reinforce topically obvious papers. Berry-LR instead selects anchors using move relevance, fused retrieval support, and query coverage, keeping feedback aligned with the survey move and target scope.

Retrieval-trajectory evidence. Finally, we show that retrieval trajectories provide crucial ranking evidence. While Direct-topic RRF only yields 0.3254 R@1500, Trajectory-RRF only applies reciprocal rank fusion over move-conditioned trajectories and reaches 0.4474. This shows that how a paper is discovered is informative beyond its original BM25 rank. The full model further reaches 0.5025 R@1500 by combining the original topic-based signal with graph evidence such as query

Variant	R@100	R@150	R@200	R@500	R@1000	R@1500
<i>Move-based search broadening</i>						
Topic-only retrieval	0.0762	0.1098	0.1396	0.1980	0.2844	0.3254
Generic query expansion	0.1171	0.1427	0.1959	0.2691	0.3573	0.4173
<i>Anchor feedback</i>						
One-round retrieval	0.1124	0.1503	0.2068	0.2953	0.3568	0.4099
Random anchors	0.1129	0.1728	0.2234	0.2997	0.3715	0.4329
Top-ranked anchors	0.1137	0.1664	0.2272	0.3054	0.3748	0.4426
<i>Graph score</i>						
Trajectory-RRF only	0.1054	0.1624	0.2396	0.3134	0.3823	0.4474
Direct-topic RRF only	0.0762	0.1098	0.1396	0.1980	0.2844	0.3254
Full Berry-LR	0.1936	0.3001	0.3341	0.4052	0.4453	0.5025

Table 4: Ablation results of different Berry-LR variants on reference retrieval. The metric is Recall@k, where k is the number of top retrieved papers. The best result in each column is shown in bold.

coverage, move coverage, round coverage, anchor participation, and agreement between direct and berrypicking retrieval. Thus, Berry-LR benefits from both staying anchored to the survey topic and rewarding papers repeatedly discovered through structured search paths.

4.4 Further Analysis

The previous experiments evaluate Berry-LR by exact recovery of human survey reference lists. This evaluation is reproducible and conservative, but a human-written reference list is only one valid reference set for a survey topic, analogous to the broader issue that retrieval evaluation often relies on incomplete relevance evidence (Buckley and Voorhees, 2004; Sakai and Kando, 2008). We therefore further analyze selected gold references, missed gold references, and selected non-gold papers to better understand what exact matching captures and what it may underestimate. Additional statistics and annotation details are provided in Appendix A.4.

Selected gold references tend to be structurally central. Gold references selected by Berry-LR tend to play more central roles in the original surveys. They appear in 2.81 citation contexts on average, compared with 1.75 for missed gold references, and are more likely to occur across multiple citation contexts and in introduction, background, or taxonomy sections. This suggests that Berry-LR preferentially recovers references that support the global framing and organization of a survey.

Missed gold references are often tied to local writing contexts. Missed gold references are often associated with local writing needs, such as a

specific subsection, dataset, benchmark, application example, or future-direction discussion. This indicates that some references become important only after the outline or local claim structure is known. Thus, the remaining misses reflect the difference between survey-level reference-set construction and section-level citation support.

Selected non-gold papers are not always false positives. Exact matching treats selected non-gold papers as false positives, but manual inspection suggests that this is overly strict. Among 300 randomly sampled selected non-gold papers, two annotators independently judged 198 as acceptable, 75 as partially acceptable, and 27 as unacceptable after resolving disagreements through discussion. Acceptable cases include plausible alternative references such as prior surveys, representative methods, or application papers, while incorrect cases mainly arise from neighboring-topic drift, task mismatch, or overestimated marginal contribution. Thus, exact reference-list recovery gives a conservative estimate of reference quality, while also highlighting the need to constrain retrieval broadening. We provide an example in Appendix D.

Conclusion

We propose Berry-LR, a literature retrieval framework for automatic scientific survey generation. Starting from a target survey topic, Berry-LR performs move-based search broadening, uses high-confidence papers as anchors for multi-round berrypicking retrieval, and records retrieval trajectories as evidence for trajectory-aware reference selection. Experiments on SurGE show that Berry-LR improves both reference retrieval and

final reference selection over strong retrieval and automatic scientific survey generation baselines. Results on SurveyBench further show that the selected reference set provides stronger reference-level support for downstream survey generation. These findings support our central motivation: survey reference-set construction requires move-based search broadening and trajectory-aware selection beyond one-shot relevance retrieval. Future work can incorporate outline-aware retrieval to better recover section-local references.

Limitations

Berry-LR is designed as a pre-writing reference selection framework, rather than a complete survey generation system. Its goal is to construct a survey-level reference set that broadly supports the target topic before the writing stage begins. This design is effective for recovering core papers with global evidential value, but it may still miss references that are needed only in more local writing contexts, such as specific subsections, section-level claims, dataset examples, benchmark discussions, or future-direction analysis. These local citation needs are often determined after the outline and detailed content have been specified. Therefore, Berry-LR can be naturally combined with existing outline-aware, section-level, or generation-time retrieval methods to provide both global reference coverage and local citation support.

Our evaluation mainly treats the references selected by human survey authors as the core literature for the corresponding survey topic, and therefore relies on exact-match reference overlap. This protocol does not imply that the author-written bibliography is the only valid reference set. There may exist other relevant and citation-worthy papers that are not included in the original survey references. To provide a complementary evaluation setting, we also evaluate Berry-LR on SurveyForge’s SurveyBench, which provides manually curated topic-level reference sets for broad domain topics. Berry-LR achieves strong performance in this setting, and our manual inspection further suggests that many selected non-gold papers are still relevant and useful. Thus, we interpret the reported overlap scores as measuring alignment with expert-selected or manually curated core literature, rather than as an exhaustive assessment of all valid references for a topic.

Finally, our main experiments are conducted

primarily on computer-science survey topics. Although these topics cover diverse subareas and retrieval scenarios, the behavior of move-based broadening, anchor feedback, and drift control may vary across disciplines with different terminology, citation practices, and corpus structures. We include a biomedical case study from SurveyGen in Appendix C, where Berry-LR also shows promising retrieval behavior. More systematic evaluation on non-computer-science survey corpora is needed to better understand the robustness of Berry-LR across disciplines.

Funding information

This work is supported by the Major Project of the National Social Science Fund of China, “Big Data-driven Semantic Evaluation System of Science and Technology Literature” (Grant No. 21&ZD329).

Acknowledgments

The AI-driven experiments, simulations, and model training were performed on the robotic AI-Scientist platform of the Chinese Academy of Sciences.

References

- Shubham Agarwal, Gaurav Sahu, Abhay Puri, Issam H. Laradji, Krishnamurthy Dj Dvijotham, Jason Stanley, Laurent Charlin, and Christopher Pal. 2025. *Litllms, llms for literature review: Are we there yet?* *Trans. Mach. Learn. Res.*, 2025.
- Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf.
- Akari Asai, Jacqueline He, Rulin Shao, Weijia Shi, Amanpreet Singh, Joseph Chee Chang, Kyle Lo, Luca Soldaini, Sergey Feldman, Mike D’Arcy, and 1 others. 2026. Synthesizing scientific literature with retrieval-augmented language models. *Nature*, pages 1–7.
- Tong Bao, Mir Tafseer Nayeem, Davood Rafiei, and Chengzhi Zhang. 2025. *Surveygen: Quality-aware scientific survey generation with large language models*. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 2712–2736. Association for Computational Linguistics.
- Marcia J Bates. 1989. The design of browsing and berrypicking techniques for the online search interface. *Online review*, 13(5):407–424.

- Chris Buckley and Ellen M. Voorhees. 2004. [Retrieval evaluation with incomplete information](#). In *SIGIR 2004: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Sheffield, UK, July 25-29, 2004, pages 25–32. ACM.
- Jaime Carbonell and Jade Goldstein. 1998. [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*, SIGIR '98, pages 335–336, New York, NY, USA. Association for Computing Machinery.
- Jing Chen, Zhiheng Yang, Yixian Shen, Jie Liu, Adam Belloum, Paola Grosso, and Chrysa Papagianni. 2025. [Surveygen-i: Consistent scientific survey generation with evolving plans and memory-guided writing](#). In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, IJCNLP-AAACL 2025, Mumbai, India, December 20-24, 2025*, pages 3687–3714. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.
- Gordon V. Cormack, Charles L. A. Clarke, and Stefan Büttcher. 2009. [Reciprocal rank fusion outperforms condorcet and individual rank learning methods](#). In *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR 2009, Boston, MA, USA, July 19-23, 2009, pages 758–759. ACM.
- Gregory Hok Tjoan Go, Khang Ly, Anders Søgaard, Seyed Amin Tabatabaei, Maarten de Rijke, and Xinyi Chen. 2026. [Lira: A multi-agent framework for reliable and readable literature review generation](#). In *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026*, pages 40456–40464. AAAI Press.
- Beichen Guo, Zhiyuan Wen, Jia Gu, Senzhang Wang, Haochen Shi, Ruosong Yang, and Shuaiqi Liu. 2026. [Surveylens: A research discipline-aware benchmark for automatic survey generation](#). *CoRR*, abs/2602.11238.
- Chao-Chun Hsu, Erin Bransom, Jenna Sparks, Bailey Kuehl, Chenhao Tan, David Wadden, Lucy Lu Wang, and Aakanksha Naik. 2024. [CHIME: llm-assisted hierarchical organization of scientific studies for literature review support](#). In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, Findings of ACL, pages 118–132. Association for Computational Linguistics.
- Yue Hu and Xiaojun Wan. 2014. [Automatic generation of related work sections in scientific papers: An optimization approach](#). In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1624–1633. ACL.
- Tetsu Kasanishi, Masaru Isonuma, Junichiro Mori, and Ichiro Sakata. 2023. [Scireviewgen: A large-scale dataset for automatic literature review generation](#). In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, Findings of ACL, pages 6695–6715. Association for Computational Linguistics.
- Yuxuan Lai, Yupeng Wu, Yidan Wang, Wenpeng Hu, and Chen Zheng. 2024. [Instruct large language models to generate scientific literature survey step by step](#). In *Natural Language Processing and Chinese Computing - 13th National CCF Conference, NLPCC 2024, Hangzhou, China, November 1-3, 2024, Proceedings, Part V*, Lecture Notes in Computer Science, pages 484–496. Springer.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. [Retrieval-augmented generation for knowledge-intensive NLP tasks](#). In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.
- Minghao Li, Ying Zeng, Zhihao Cheng, Cong Ma, and Kai Jia. 2025. [Reportbench: Evaluating deep research agents via academic survey tasks](#). *CoRR*, abs/2508.15804.
- Xiangci Li, Biswadip Mandal, and Jessica Ouyang. 2022. [CORWA: A citation-oriented related work annotation dataset](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 5426–5440. Association for Computational Linguistics.
- Maria Liakata. 2010. [Zones of conceptualisation in scientific papers: a window to negative and speculative statements](#). In *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing, NeSp-NLP@ACL 2010, Uppsala, Sweden, July 10, 2010*, pages 1–4. University of Antwerp.
- Xun Liang, Jiawei Yang, Yezhaohui Wang, Chen Tang, Zifan Zheng, Shichao Song, Zehao Lin, Yebin Yang, Simin Niu, Hanyu Wang, Bo Tang, Feiyu Xiong, Keming Mao, and Zhiyu Li. 2025. [Surveyx: Academic survey automation via large language models](#). *CoRR*, abs/2502.14776.
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. [G-eval: NLG evaluation using gpt-4 with better human alignment](#). In *Proceedings of the 2023 Conference on*

- Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 2511–2522. Association for Computational Linguistics.
- Yinqi Liu, Yueqi Zhu, Yongkang Zhang, Xinfeng Li, Feiran Liu, Yufei Sun, Xin Wang, Renzhao Liang, Yidong Wang, and Cunxiang Wang. 2026. [MVSS: A unified framework for multi-view structured survey generation](#). *CoRR*, abs/2601.09504.
- Yao Lu, Yue Dong, and Laurent Charlin. 2020. [Multixscience: A large-scale dataset for extreme multi-document summarization of scientific articles](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 8068–8074. Association for Computational Linguistics.
- Zhaoyu Ma, Yuan Shan, Jiahao Zhao, Nan Xu, and Lei Wang. 2025. [Meow: End-to-end outline writing for automatic academic survey](#). *CoRR*, abs/2509.19370.
- Minh-Anh Nguye, Minh-Duc Nguyen, Ha Lan N. T., Hai-Dang Kieu, Nguyen Tien Dong, and Dung D. Le. 2025. [Surveyg: A multi-agent LLM framework with hierarchical citation graph for automated survey generation](#). *CoRR*, abs/2510.07733.
- Kazuya Nishimura, Kuniaki Saito, Tosho Hirasawa, and Yoshitaka Ushiku. 2024. Toward structured related work generation with novelty statements. In *Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024)*, pages 38–57.
- OpenAI. 2024. [Gpt-4 technical report](#). *Preprint*, arXiv:2303.08774.
- Stephen E. Robertson and Hugo Zaragoza. 2009. [The probabilistic relevance framework: BM25 and beyond](#). *Found. Trends Inf. Retr.*, 3(4):333–389.
- Tetsuya Sakai and Noriko Kando. 2008. [On information retrieval metrics designed for evaluation with incomplete relevance assessments](#). volume 11, pages 447–470.
- Yijia Shao, Yucheng Jiang, Theodore A. Kanell, Peter Xu, Omar Khattab, and Monica S. Lam. 2024. [Assisting in writing wikipedia-like articles from scratch with large language models](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 6252–6278. Association for Computational Linguistics.
- Xiaofeng Shi, Qian Kou, Yuduo Li, Ning Tang, Jinxin Xie, Longbin Yu, Songjing Wang, and Hua Zhou. 2025. [Scisage: A multi-agent framework for high-quality scientific survey generation](#). *CoRR*, abs/2506.12689.
- Zhengliang Shi, Shen Gao, Zhen Zhang, Xiuying Chen, Zhumin Chen, Pengjie Ren, and Zhaochun Ren. 2023. [Towards a unified framework for reference retrieval and related work generation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, Findings of ACL, pages 5785–5799. Association for Computational Linguistics.
- Amanpreet Singh, Mike D’Arcy, Arman Cohan, Doug Downey, and Sergey Feldman. 2023. [SciRepeval: A multi-format benchmark for scientific document representations](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 5548–5566. Association for Computational Linguistics.
- Hannah Snyder. 2019. Literature review as a research methodology: An overview and guidelines. *Journal of business research*, 104:333–339.
- Weihang Su, Anzhe Xie, Qingyao Ai, Jianming Long, Xuanyi Chen, Jiaxin Mao, Ziyi Ye, and Yiqun Liu. 2025. [Surge: A benchmark and evaluation framework for scientific survey generation](#). *arXiv preprint arXiv:2508.15658*.
- Xuemei Tang, Xufeng Duan, and Zhenguang G. Cai. 2025. [Large language models for automated literature review: An evaluation of reference generation, abstract writing, and review composition](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 1602–1617. Association for Computational Linguistics.
- Qwen Team. 2026. [Qwen3. 5-omni technical report](#). *arXiv preprint arXiv:2604.15804*.
- Simone Teufel. 1999. *Argumentative zoning: Information extraction from scientific text*. Ph.D. thesis, University of Edinburgh Edinburgh, Scotland.
- Richard J Torraco. 2005. Writing integrative literature reviews: Guidelines and examples. *Human resource development review*, 4(3):356–367.
- Yidong Wang, Qi Guo, Wenjin Yao, Hongbo Zhang, Xin Zhang, Zhen Wu, Meishan Zhang, Xinyu Dai, Min Zhang, Qingsong Wen, Wei Ye, Shikun Zhang, and Yue Zhang. 2024. [Autosurvey: Large language models can automatically write surveys](#). In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.
- Zhiyuan Wen, Jiannong Cao, Zian Wang, Beichen Guo, Ruosong Yang, and Shuaiqi Liu. 2025. [Interactivesurvey: An llm-based personalized and interactive survey paper generation system](#). *CoRR*, abs/2504.08762.
- Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muenighoff, Defu Lian, and Jian-Yun Nie. 2024. [C-pack:](#)

Packed resources for general chinese embeddings. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024*, pages 641–649. ACM.

Xiangchao Yan, Shiyang Feng, Jiakang Yuan, Renqiu Xia, Bin Wang, Lei Bai, and Bo Zhang. 2025. **SURVEYFORGE : On the outline heuristics, memory-driven generation, and multi-dimensional evaluation for automated survey writing**. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 12444–12465. Association for Computational Linguistics.

Hongbo Zhang, Han Cui, Yidong Wang, Yijian Tian, Qi Guo, Cunxiang Wang, Jian Wu, Chiyu Song, and Yue Zhang. 2025. **Deep literature survey automation with an iterative workflow**. *CoRR*, abs/2510.21900.

Ming Zhang, Jiabao Zhuang, Wenqing Jing, Ziyu Kong, Jingyi Deng, Yujiong Shen, Kexin Tan, Yuhang Zhao, Ning Luo, Renzhe Zheng, and 1 others. 2026. Can deep research agents find and organize? evaluating the synthesis gap with expert taxonomies. *arXiv preprint arXiv:2601.12369*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. **Judging llm-as-a-judge with mt-bench and chatbot arena**. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Kun Zhu, Xiaocheng Feng, Xiachong Feng, Yingsheng Wu, and Bing Qin. 2023. **Hierarchical catalogue generation for literature review: A benchmark**. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6790–6804, Singapore. Association for Computational Linguistics.

A Additional Experimental Details

A.1 Evaluation Protocol and Metric Details

This appendix provides additional details about the datasets, evaluation protocols, baselines, and budget control used in our experiments. The main paper summarizes the high-level settings; here we further specify how SurGE and SurveyBench are adapted to evaluate literature retrieval and reference-set construction.

Datasets. We evaluate Berry-LR on SurGE and SurveyBench. SurGE contains 205 verified surveys and a retrieval corpus of 1,086,992 candidate papers (Su et al., 2025). For each survey, we use its reference list as the gold reference set. This

provides a conservative and reproducible signal for evaluating whether a retrieval method can recover papers cited by expert-written surveys. SurveyForge’s SurveyBench provides a complementary topic-level setting for evaluating whether selected references can support downstream survey generation (Yan et al., 2025).

SurGE retrieval protocol. SurGE formulates automatic scientific survey generation as a two-stage process: retrieving a relevant paper set from a large academic corpus and generating a structured survey from the retrieved evidence (Su et al., 2025). Our SurGE retrieval experiment focuses on the first stage. For each survey topic q , let G_q denote the gold reference set extracted from the expert-written survey, and let π_q be the ranked list returned by a retrieval method. The top- k retrieved set is denoted as $\pi_q^{(k)}$. We compute reference recall at budget k as:

$$R@k(q) = \frac{|\pi_q^{(k)} \cap G_q|}{|G_q|}.$$

The final score is averaged over all SurGE surveys:

$$R@k = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} R@k(q).$$

In SurGE, Berry-LR starts from the survey title and expands retrieval internally through survey moves and anchor feedback. We therefore compare baseline retrievers under both title-only and title+abstract query settings. We report two Berry-LR variants: Berry-LR (BM25) and Berry-LR (RoBERTa retriever). Unless otherwise specified, Berry-LR refers to the BM25-based version.

SurGE Reference Selection. Retrieval recall measures whether gold references appear in the candidate pool, while survey writing also requires selecting a compact reference set from retrieved candidates. For reference selection, each method outputs a selected set S_q for survey topic q . We evaluate it against the same gold reference set G_q using:

$$\text{Precision}(q) = \frac{|S_q \cap G_q|}{|S_q|},$$

$$\text{Recall}(q) = \frac{|S_q \cap G_q|}{|G_q|},$$

$$F1(q) = \frac{2 \text{Precision}(q) \text{Recall}(q)}{\text{Precision}(q) + \text{Recall}(q)}.$$

We report macro-averaged precision, recall, and F1 across all survey topics.

To keep generation and selection comparisons controlled while preserving retrieval baselines’ native rankings, we separate the protocols. Retrieval methods that return ranked lists are converted into selected reference sets by taking their top 50 papers from their own ranked outputs. Automatic scientific survey generation baselines are evaluated using the references actually cited in their generated surveys, with Qwen-3.5-Plus used as the generation model and the Berry-LR (BM25) candidate pool supplied as controlled input (Team, 2026). Berry-LR is evaluated on the same controlled candidate pool using the reference set produced by trajectory-aware reference selection.

Relation to SurGE generation metrics. SurGE also defines end-to-end survey generation metrics, including citation accuracy, structural quality, and content quality. Citation accuracy is evaluated at document, section, and sentence levels, while structural and content quality are evaluated over the generated survey (Su et al., 2025). These metrics are designed for complete survey generation. Since Berry-LR focuses on literature retrieval and reference-set construction, our SurGE experiments use reference retrieval and reference selection as the primary evaluation targets.

SurveyBench reference-quality protocol. For downstream survey generation evaluation, we follow the reference-quality protocol used by SurveyForge on SurveyBench (Yan et al., 2025). SurveyForge evaluates generated surveys with Survey Assessment Metrics (SAM), including reference quality, outline quality, and content quality. Since Berry-LR focuses on literature retrieval and reference selection, we report the reference-quality metrics: Input Coverage and Reference Coverage.

For downstream reference-quality evaluation, we follow the SurveyForge evaluation protocol on SurveyBench (Yan et al., 2025). Input Coverage evaluates the overlap between the literature supplied to the survey generator and the benchmark references, while Reference Coverage evaluates the benchmark alignment of the references cited in the generated survey. We compute both metrics using the same evaluation procedure across all compared methods.

Budget control on SurveyBench. Input Coverage and Reference Coverage evaluate different stages and therefore use different budgets. For Input Coverage, we evaluate the retrieved input literature under an estimated budget of 700 papers,

approximating the scale of the input literature used by SurveyForge. This metric reflects the raw coverage of the retrieval pool before reference selection. For Reference Coverage, Berry-LR first selects a compact final reference set from the retrieved candidates. To avoid improving coverage simply by selecting more papers, we retain the same number of final references as SurveyForge for each topic before survey generation. We report results under the same generator settings used in the main experiments.

Baselines. For SurGE retrieval, we compare with BM25 (Robertson and Zaragoza, 2009), BGE (Xiao et al., 2024), SPECTER2 (Singh et al., 2023), OpenScholar retriever (Asai et al., 2026), and the RoBERTa retriever from SurGE (Su et al., 2025). For SurGE reference selection and SurveyBench evaluation, we compare with retrieval-based selectors and representative automatic scientific survey generation systems, including AutoSurvey (Wang et al., 2024), SurveyForge (Yan et al., 2025), and SurveyGen (Bao et al., 2025).

Interpretation of exact-match reference metrics. Both SurGE and SurveyBench rely on overlap with expert or benchmark references (Su et al., 2025; Yan et al., 2025). These exact-match metrics are reproducible and useful for comparing systems, but they are conservative. A human-written reference list is not the only valid reference set for a survey topic, and incomplete relevance assessments may underestimate valid alternatives in retrieval-style evaluation (Buckley and Voorhees, 2004; Sakai and Kando, 2008). We therefore interpret reference overlap as alignment with expert-selected core literature rather than as an exhaustive measure of all useful references. This motivates our additional analysis of selected gold references, missed gold references, and selected non-gold papers.

A.2 Additional Method Details

This appendix provides implementation details omitted from the main method section for compactness. These details concern anchor scoring, drift control, and trajectory-aware reference selection.

Implementation details. Unless otherwise stated, we keep the retrieval and selection hyperparameters fixed across datasets and ablation variants; evaluation budgets follow the dataset-specific protocols described above. Berry-LR uses the six survey moves defined in the main method

Hyperparameter	Value
Survey moves	6
Initial queries per move	4
Expansion queries per move	4
Maximum retrieval rounds	3
Anchors per move-round	5
Local candidates per move-round	20
BM25 candidates per query	250
Dense candidates per query	1000
RRF constant k_{RRF}	60
Round decay β	0.70
Stopping starts from round	2

Table 5: Main retrieval hyperparameters used by Berry-LR. The dense retrieval setting is used for SurveyBench, while the BM25 retrieval setting is used for SurGE unless otherwise specified.

section: background, purpose, methods, results, limitations, and conclusion. For each move, the system generates four initial queries and four expansion queries per active round. We run at most three berrypicking rounds and select up to five anchors for each move at each round. Retrieved lists are fused with reciprocal rank fusion using $k_{\text{RRF}} = 60$. On SurGE, each retrieval query retrieves 250 papers under the BM25 setting. On SurveyBench, each dense query retrieves 1000 papers; the reported Input Coverage still follows the budget control described in Appendix A.1.

Anchor scoring and stopping. For each move m and retrieval round r , Berry-LR first forms a fused candidate list $\mathcal{C}_m^{(r)}$ by merging the retrieval results for the move-conditioned queries. It then scores each candidate paper $d \in \mathcal{C}_m^{(r)}$ by:

$$a_m^{(r)}(d) = \lambda_1 \text{sim}_m(d) + \lambda_2 \tilde{s}_m^{(r)}(d) + \lambda_3 \text{cov}_m^{(r)}(d) + \lambda_4 \text{titleSim}_x(d).$$

Here, $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ are non-negative weights. We set $(\lambda_1, \lambda_2, \lambda_3, \lambda_4) = (0.56, 0.24, 0.16, 0.04)$, corresponding to move relevance, normalized fused retrieval support, query coverage, and lexical title similarity. The term $\text{sim}_m(d)$ measures the similarity between paper d and the current survey move. We instantiate each move as a lexical term set and compute this score by weighted overlap with the title and abstract of d . The term $\tilde{s}_m^{(r)}(d)$ is the normalized RRF score under move m in round r . The term $\text{cov}_m^{(r)}(d)$ measures the fraction of queries under move m that retrieve d :

$$\text{cov}_m^{(r)}(d) = \frac{|\{q \in \mathcal{Q}_m^{(r)} : d \in R(q)\}|}{|\mathcal{Q}_m^{(r)}|}.$$

Finally, $\text{titleSim}_x(d)$ measures lexical similarity between the target survey topic and the title of d . Only candidates whose anchor score exceeds a threshold are eligible to become anchors:

$$\mathcal{A}_m^{(r)} = \{d \in \mathcal{C}_m^{(r)} : a_m^{(r)}(d) \geq \tau_{\text{anchor}}\}.$$

The temporary anchor set is the top- L subset of this eligible pool:

$$\tilde{\mathcal{H}}_m^{(r)} = \text{TopL}_{d \in \mathcal{A}_m^{(r)}} a_m^{(r)}(d).$$

If $\mathcal{A}_m^{(r)} = \emptyset$, then $\mathcal{H}_m^{(r)} = \emptyset$ and Berry-LR stops expanding move m . If $r = 1$, the temporary anchors are used directly for subsequent query expansion. Starting from $r = 2$, they are passed to drift control before use.

Drift control. Anchor feedback can broaden retrieval, but it may also move the search toward neighboring topics. Berry-LR therefore applies a move-level drift-control rule. This check uses vector representations only for measuring topical drift, while the anchor relevance score above uses lexical overlap and retrieval-support features.

Let $\mathbf{u}_m$ be the vector representation of move m , constructed from the target survey topic and the initial move-conditioned queries. Let $\mathbf{h}_d$ denote the representation of paper d . Starting from the second retrieval round, Berry-LR checks the temporary anchors $\tilde{\mathcal{H}}_m^{(r)}$ against the initial move-conditioned search state. For this temporary anchor set, we compute the average and maximum anchor alignment:

$$\mu_m^{(r)} = \frac{1}{|\tilde{\mathcal{H}}_m^{(r)}|} \sum_{d \in \tilde{\mathcal{H}}_m^{(r)}} \cos(\mathbf{h}_d, \mathbf{u}_m),$$

$$\nu_m^{(r)} = \max_{d \in \tilde{\mathcal{H}}_m^{(r)}} \cos(\mathbf{h}_d, \mathbf{u}_m).$$

The drift-control indicator is:

$$\text{stop}(m, r) = \mathbb{I}[\mu_m^{(r)} < \max(\tau_\mu, \rho_\mu \mu_m^{(1)}) \wedge \nu_m^{(r)} < \max(\tau_\nu, \rho_\nu \nu_m^{(1)})].$$

Here, τ_μ and τ_ν are absolute thresholds, while ρ_μ and ρ_ν define relative thresholds against the first-round anchor similarities. For topical query batches, we use $(\tau_\mu, \tau_\nu, \rho_\mu, \rho_\nu) = (0.220, 0.240, 0.78, 0.70)$. For supporting-reference query batches, we use relaxed thresholds $(0.120, 0.145, 0.52, 0.42)$. If $\text{stop}(m, r) = 1$, we set $\mathcal{H}_m^{(r)} = \emptyset$ and stop

expanding move m . Otherwise, $\mathcal{H}_m^{(r)} = \tilde{\mathcal{H}}_m^{(r)}$ and the selected anchors are used to generate the next-round queries. This rule allows anchor-based exploration while preventing uncontrolled topic drift.

Sequential selection details. Trajectory-aware reference selection combines LLM-based set-conditioned judgments with a trajectory-derived prior. For candidate $\hat{d}_i$, the LLM estimates topical relevance, citation value, and marginal gain:

$$\mathbf{z}_i = [z_{\text{rel}}(\hat{d}_i), z_{\text{val}}(\hat{d}_i), z_{\text{gain}}(\hat{d}_i | \mathcal{S}_{i-1})].$$

The three scores are produced as 0–10 numeric judgments, clamped when necessary, and normalized to $[0, 1]$. Similar LLM-as-judge protocols have been widely used for evaluating long-form generation when direct automatic metrics are insufficient (Liu et al., 2023; Zheng et al., 2023; Su et al., 2025).

The LLM score is derived from the normalized decision signals:

$$s_{\text{LLM}}(\hat{d}_i | \mathcal{S}_{i-1}) = \eta_1 \bar{z}_{\text{rel}}(\hat{d}_i) + \eta_2 \bar{z}_{\text{val}}(\hat{d}_i) + \eta_3 \bar{z}_{\text{gain}}(\hat{d}_i | \mathcal{S}_{i-1}).$$

where η_1, η_2, η_3 are non-negative weights. The trajectory-derived prior score is computed from the normalized citation-worthiness score and the trajectory-derived prior rank:

$$s_{\text{graph}}(\hat{d}_i) = \delta_1 \bar{c}(\hat{d}_i) + \delta_2 \left(1 - \frac{\text{rank}_{\mathcal{C}}(\hat{d}_i) - 1}{N - 1} \right),$$

where $\bar{c}(\hat{d}_i)$ is the normalized citation-worthiness score and $\text{rank}_{\mathcal{C}}(\hat{d}_i)$ is the rank of $\hat{d}_i$ in the trajectory-derived prior ranking. The final selection score is:

$$\pi_i = \alpha s_{\text{LLM}}(\hat{d}_i | \mathcal{S}_{i-1}) + (1 - \alpha) s_{\text{graph}}(\hat{d}_i).$$

Here, α controls the balance between set-conditioned LLM judgment and retrieval-trajectory graph evidence. The candidate is accepted if $\pi_i \geq \gamma$ and the selected set has fewer than K papers.

A.3 Evidence Features for Citation-worthiness

This section provides the detailed feature construction used in the trajectory evidence graph. A retrieval path is defined as a tuple

$$p = (m, r, q),$$

Selection hyperparameter	Value
Candidate window	150
LLM weight α	0.70
Trajectory-derived prior weight $1 - \alpha$	0.30
Acceptance threshold γ	0.60

Table 6: Hyperparameters for trajectory-aware reference selection.

where m is a survey move, r is the retrieval round, and $q \in \mathcal{Q}_m^{(r)}$ is the query that retrieves the paper. For each candidate paper d , let $\mathcal{P}(d)$ denote the set of retrieval paths that reach d .

Retrieval-strength evidence. The retrieval-strength group $\mathbf{e}_{\text{ret}}(d)$ measures how strongly d is supported by both berrypicking retrieval and direct topic-based retrieval:

$$\mathbf{e}_{\text{ret}}(d) = [e_{\text{graph}}(d), e_{\text{base}}(d), e_{\text{rank}}(d)].$$

The graph retrieval score aggregates reciprocal-rank evidence over all move–round–query paths:

$$\tilde{e}_{\text{graph}}(d) = \sum_{(m,r,q) \in \mathcal{P}(d)} \beta^{r-1} \frac{1}{k + \text{rank}_q(d)},$$

where $\beta \in (0, 1]$ discounts later retrieval rounds and k is the reciprocal-rank smoothing constant. We then max-normalize it over all candidates:

$$e_{\text{graph}}(d) = \frac{\tilde{e}_{\text{graph}}(d)}{\max_{d' \in \mathcal{C}} \tilde{e}_{\text{graph}}(d')}.$$

The direct baseline score $e_{\text{base}}(d)$ is computed from retrieval using the target survey specification as the query. The best-rank feature records the strongest single-path rank signal:

$$e_{\text{rank}}(d) = \frac{1}{1 + \min_{(m,r,q) \in \mathcal{P}(d)} \text{rank}_q(d)}.$$

Trajectory evidence. The trajectory group $\mathbf{e}_{\text{traj}}(d)$ captures how broadly d is discovered:

$$\mathbf{e}_{\text{traj}}(d) = [e_{\text{query}}(d), e_{\text{move}}(d), e_{\text{round}}(d), e_{\text{path}}(d)].$$

Here, $e_{\text{query}}(d)$ counts the number of distinct queries that retrieve d , $e_{\text{move}}(d)$ counts the number of survey moves under which d appears, and $e_{\text{round}}(d)$ counts the number of retrieval rounds in which d is reached. The path-diversity feature is defined as the number of distinct retrieval paths:

$$e_{\text{path}}(d) = |\mathcal{P}(d)|.$$

In implementation, count features are log-normalized:

$$\text{lognorm}(x) = \frac{\log(1+x)}{\max_{d' \in \mathcal{C}} \log(1+x_{d'})}.$$

Anchor evidence. The anchor-evidence group $\mathbf{e}_{\text{anc}}(d)$ records whether a paper actively guides the berrypicking process:

$$\mathbf{e}_{\text{anc}}(d) = [e_{\text{is_anchor}}(d), e_{\text{expand_count}}(d)].$$

The first feature indicates whether d is selected as an anchor in any round:

$$e_{\text{is_anchor}}(d) = \mathbb{I} \left[d \in \bigcup_{m,r} \mathcal{H}_m^{(r)} \right].$$

The second feature counts how many later queries are generated from an anchor set containing d :

$$e_{\text{expand_count}}(d) = \left| \left\{ (m, r, q') : q' \in \mathcal{Q}_m^{(r+1)}, d \in \mathcal{H}_m^{(r)} \right\} \right|.$$

Agreement evidence. The agreement group $\mathbf{e}_{\text{agree}}(d)$ captures whether different retrieval sources support the same candidate:

$$\mathbf{e}_{\text{agree}}(d) = [e_{\text{overlap}}(d), e_{\text{graph-base}}(d)].$$

The overlap feature indicates whether d appears in both berrypicking retrieval and the direct baseline:

$$e_{\text{overlap}}(d) = \mathbb{I}[e_{\text{graph}}(d) > 0 \wedge e_{\text{base}}(d) > 0].$$

We also include a soft agreement score:

$$e_{\text{graph-base}}(d) = e_{\text{graph}}(d) \cdot e_{\text{base}}(d),$$

which assigns higher values to papers strongly supported by both sources.

Citation-worthiness score. All evidence features are normalized and concatenated into $\phi(d)$. The citation-worthiness score is computed as:

$$c(d) = \mathbf{w}^\top \phi(d).$$

The resulting score defines the trajectory-derived prior used in sequential reference selection.

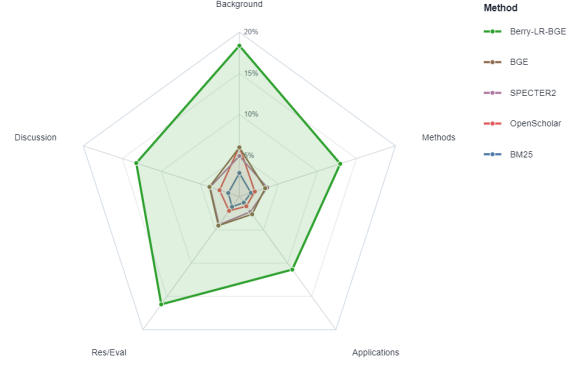


Figure 3: Section-wise coverage of gold references recovered by different retrieval methods. Each axis measures the fraction of gold references cited in a given body-section type that are recovered by the method. Compared with one-shot retrieval baselines, Berry-LR (BM25) covers gold references across a broader and more balanced set of survey-writing contexts.

A.4 Further Analysis Details

This appendix provides additional details for the further analysis in Section 4.4. The goal of this analysis is to clarify how exact reference-list matching should be interpreted. Although exact matching is reproducible and useful for comparing systems, a survey reference list is not an exhaustive set of all useful references for a topic (Buckley and Voorhees, 2004; Sakai and Kando, 2008). We therefore analyze three complementary cases: gold references selected by Berry-LR, gold references missed by Berry-LR, and selected papers that do not appear in the gold reference set.

Figure 3 gives a complementary view of exact-match recovery from the perspective of survey writing structure. Instead of asking only how many gold references are recovered overall, the figure asks whether the recovered gold references cover the different parts of a human-written survey. Each dimension corresponds to a body-section type, such as background, methods, applications, resources/evaluation, and discussion. A more complete and regular shape therefore indicates that a method recovers gold references across multiple survey-writing roles rather than concentrating on a narrow part of the article.

The comparison shows that one-shot retrieval baselines tend to produce more uneven coverage profiles. They recover some gold references that are directly aligned with the topic query, but their coverage is less consistent across different section types. In contrast, Berry-LR (BM25) produces a fuller and more balanced profile. This suggests that

the berrypicking process does not merely increase exact-match recovery, but also helps recover references used by human survey writers in a wider range of rhetorical and organizational contexts.

Selected gold references. We first examine gold references that are successfully selected by Berry-LR. Using their citation contexts in the original surveys, we compare selected gold references with missed gold references. Selected gold references appear in 2.81 citation contexts on average, compared with 1.75 for missed gold references. They also have a higher multi-context ratio, 64.26% versus 37.48%. In addition, selected gold references are more likely to appear in introduction, background, or taxonomy sections, whereas missed references are more often confined to fine-grained subsections.

These patterns suggest that Berry-LR tends to recover references that play central organizational roles in human-written surveys. Such papers often define the field, motivate the survey topic, support major methodological categories, or help structure the taxonomy of the survey. Therefore, the recovered gold references are not arbitrary exact matches. They are often the papers that survey writers use to establish the global framing of a topic.

Missed gold references. We then analyze gold references that are not selected by Berry-LR. Manual inspection suggests that many missed references depend on local writing contexts. They often support a specific subsection, dataset, benchmark, application example, or future-direction discussion. These papers may be important for a particular sentence or section, but they are less likely to be repeatedly cited across the survey or to define the survey-level structure.

This failure mode is qualitatively different from the global literature-discovery problem targeted by Berry-LR. Berry-LR starts from a target survey topic and is designed to construct a globally useful reference set before writing begins. In contrast, some gold references become relevant only after the outline, section plan, or local claim is known. This suggests that Berry-LR can complement existing automatic scientific survey generation pipelines: it can provide the global reference set, while outline-aware retrieval or citation-context prediction can later retrieve section-local references for specific examples, datasets, claims, or future directions.

Selected non-gold papers. Finally, we examine papers selected by Berry-LR that do not appear in the gold reference set. Exact matching counts all such papers as false positives. However, this assumption can be overly strict because different surveys on the same topic may reasonably cite different references, especially when the selected papers cover broader, newer, or alternative directions in the field.

In a manual inspection of 300 selected non-gold samples, 198 are judged acceptable, 75 partially acceptable, and 27 clearly unacceptable, accounting for 66.0%, 25.0%, and 9.0%, respectively. We use the following annotation criteria. A paper is acceptable if it is a prior survey, representative method, resource, or application paper that could plausibly support the survey topic. It is partially acceptable if it is related but differs in emphasis, scope, or specificity. It is unacceptable if it mainly reflects neighboring-topic drift, task mismatch, or overestimated marginal contribution.

These findings clarify the interpretation of Berry-LR's precision. Some selected non-gold papers are genuine errors, but others are plausible references penalized only because they do not match the particular reference list written by the original survey authors. Thus, exact reference-list recovery gives a conservative estimate of reference quality. At the same time, the analysis highlights the main trade-off in Berry-LR: broadening retrieval helps discover useful literature beyond the initial topic wording, but the expanded search space must be bounded to avoid drifting into neighboring topics.

B Prompt Templates

Berry-LR is evaluated in the title-only setting: the external input is only the target survey title. The system first asks the language model to infer a compact working abstract from the title. This generated text is used only as an internal aid for query planning and drift control; it is not treated as gold metadata. The remaining prompts use this generated abstract to construct move-conditioned retrieval queries, perform anchor-based berrypicking expansion, and select a compact final source set.

B.1 Title-Conditioned Working Abstract

INPUT. target survey title.

OUTPUT. a JSON object containing a short working_abstract, salient key_terms, and brief scope_notes.

USE. the generated abstract provides a lightweight survey specification for later retrieval prompts.

Prompt B.1: Working abstract generation

You are an academic literature-retrieval assistant. Given only the title of a target survey paper, infer a compact working abstract of the survey scope. The abstract will be used only for retrieval query planning. Do not invent specific claims, results, or citations.

Requirements. Stay faithful to the survey title. Expand the title into a concise description of the likely research area, main problems, common methods, resources, evaluation settings, and neighboring areas. Include useful search vocabulary such as task names, method families, datasets or benchmarks when they are commonly associated with the topic, and application domains. Mark uncertain terms as possible rather than factual. Avoid over-specializing the topic beyond what the title supports. Return JSON only, no markdown.

Schema. working_abstract: short paragraph; key_terms: list of search terms; scope_notes: in-scope and avoid notes.

Survey title: <TARGET_SURVEY_TITLE>

B.2 Move-Conditioned Initial Queries

INPUT. target survey title, generated working abstract, and number of queries per move.

OUTPUT. a JSON object containing move_analysis and move_queries.

USE. the queries initialize six retrieval branches corresponding to survey-writing moves.

Prompt B.2: Initial query generation

You are an academic literature retrieval expert. Given a target survey title and a generated working abstract, split the retrieval intent into six survey-writing moves and generate initial retrieval queries for each move.

Moves. *background:* research setting, core domain, foundational concepts, and central topic; *purpose:* survey scope, task framing, objective, and coverage boundary; *methods:* algorithms, models, architectures, frameworks, and technical routes; *results:* datasets, benchmarks, metrics, empirical comparisons, and evaluation protocols; *limitations:* bottlenecks, failure modes, open problems, and unresolved issues; *conclusion:* synthesized takeaways, representative predecessors, and forward-looking directions.

Coverage guidance. Keep the six moves as the only retrieval branches. Within these moves, make sure the queries can retrieve not only papers that exactly repeat the title words, but also papers that a survey would normally need for context, methods, datasets, tools, evaluation, applications, or prior organization of the field. Do not create separate categories for these citation needs; blend them naturally into the relevant move queries.

Requirements. Fill every move_analysis field with a short English phrase or sentence. Generate exactly <QUERIES_PER_MOVE> concise English retrieval queries for each move. Preserve concrete terms, named methods, datasets, benchmarks, tool names, task names, and acronyms when suggested by the title or working abstract. Include a few queries that may reach important supporting

papers whose titles do not exactly repeat the survey title, while still remaining useful for the target survey. Do not use Boolean operators such as AND/OR. Return JSON only, no markdown.

Inputs. Survey title: <TARGET_SURVEY_TITLE>
Generated working abstract:
<GENERATED_WORKING_ABSTRACT>

B.3 Anchor-Based Berrypicking Expansion

INPUT. target survey title, generated working abstract, current round, move analysis, previous queries, and selected anchors for each active move.

OUTPUT. a JSON object containing next-round move_queries and short notes.

USE. anchors provide new vocabulary and concepts for the next retrieval round while the prompt constrains topical drift.

Prompt B.3: Anchor-based expansion

You are simulating berrypicking-style academic literature search. A researcher is searching for source papers for a title-only target survey. For each active survey-writing move below, use the newly selected anchor papers to reformulate the next round of retrieval queries.

Important. These anchors were selected from retrieval results by relevance and graph evidence, not from gold citations. Continue from each move's local concept neighborhood, but add useful nearby vocabulary that may discover missing papers. Useful survey references may include background papers, representative methods, datasets, benchmarks, tools, evaluation protocols, applications, or prior surveys. Use these only as soft coverage cues, not as separate branches. Do not drift into topics that are not useful for writing the target survey. Preserve named methods, datasets, benchmarks, tools, acronyms, tasks, and theory vocabulary from the anchors when they remain relevant. It is acceptable for a query to omit some exact survey-title words if the query follows a useful anchor concept and remains within the target survey scope. Do not use Boolean operators such as AND/OR. Return JSON only, no markdown.

Requirement. Generate exactly <EXPANSION_QUERIES_PER_MOVE> queries for each active move.

Inputs. Survey title: <TARGET_SURVEY_TITLE>
Generated working abstract:
<GENERATED_WORKING_ABSTRACT>
Round: <ROUND_NUMBER>
Active move branches:
<ACTIVE_MOVE_BRANCHES_JSON>

Each active branch contains the move analysis, previous queries, selected anchor papers, and retrieval-batch statistics. For example, an active *methods* branch contains a short move analysis, previous method-oriented queries, anchor titles with metadata, and a batch-similarity score.

B.4 Trajectory-aware Sequential Selection

INPUT. target survey title, generated working abstract, current selected reference titles, and one candidate paper title.

OUTPUT. a JSON object containing only `topic_relevance`, `paper_value`, and `marginal_gain`.

USE. the selector estimates whether the candidate adds useful marginal coverage to the current source set.

Prompt B.4: Sequential reference selection

You are an expert academic literature-review assistant building a final bibliography for a survey paper.

Task. Judge whether one candidate paper should be included in the final source literature set of the target survey. The external input to the system is only the survey title. The working abstract below was generated from the title and is used only to clarify the intended scope. The candidate has already been retrieved by a multi-round graph/BM25 process, so treat it as a plausible citation candidate unless the title is clearly irrelevant. Because the candidate input is only a title, avoid rejecting useful survey references merely because the title is indirect or does not repeat the exact survey title words.

Scores. Assign 0–10 scores for: `topic_relevance`, whether the candidate is useful for the survey topic; `paper_value`, whether the candidate appears useful to cite in a survey bibliography; `marginal_gain`, whether adding this candidate improves the current selected set.

Inclusion policy. This is a survey bibliography, not a tiny list of only the most exact papers. A good final set should be compact but broad enough to support background, methods, resources, evaluation, applications, and prior organization when relevant. Be moderately permissive when a paper is plausibly useful for explaining the survey area, comparing methods, giving background, or covering a missing part of the current selected set. Do not require exact keyword overlap with the survey title. Give high `marginal_gain` when a candidate fills a missing part of the current source set, even if the title is only indirectly related. Return JSON only with the three score fields above, no markdown.

Inputs. Survey title: <TARGET_SURVEY_TITLE>
Generated working abstract:
<GENERATED_WORKING_ABSTRACT>
Currently selected reference titles:
<SELECTED_TITLES>
Candidate reference title:
<CANDIDATE_REFERENCE_TITLE>

Selection schema. The sequential selector returns a structured record with the following fields: `topic_relevance`, `paper_value`, and `marginal_gain`. These are the only LLM-produced atomic judgments; the final selection score is computed deterministically from these normalized judgments and the trajectory-derived prior.

C Biomedical Case Study from SurveyGen

To test whether Berry-LR remains useful outside computer-science survey topics, we inspected a biomedical SurveyGen case: *The effectiveness of pressure therapy (15–25 mmHg) for hypertrophic burn scars: A systematic review and meta-analysis*. The survey is a 2017 dermatology meta-analysis on pressure therapy for hypertrophic scars after burn injury. Its gold bibliography contains 49 references. We follow the same title-only external-validation protocol used for SurveyGen: the only external input to both direct retrieval and Berry-LR is the target survey title. Berry-LR may generate a working abstract internally from the title for query planning and drift control, but it does not use the gold abstract or any gold metadata. It expands the topic with survey-aware retrieval moves, applies anchor-guided berrypicking with its stopping criterion, reranks candidates using graph evidence, and then performs sequential reference selection from the top-150 candidate window with a 50-paper output budget.

Retrieval behavior. The generated retrieval plan is clinically interpretable rather than being a flat lexical expansion of the title. Some queries emphasize hypertrophic burn-scar pathophysiology, pressure therapy, neighboring keloid management, and prior scar-management reviews. Others focus on randomized controlled trials, the 15–25 mmHg intervention scope, meta-analysis design, Vancouver Scar Scale outcomes, scar thickness, redness, pigmentation, and hardness. The resulting anchors include prior evidence syntheses and intervention studies such as *Pressure therapy for the control of hypertrophic scar formation after burn injury. A history and review, The effectiveness of pressure garment therapy for the prevention of abnormal scarring after burn injury: a meta-analysis*, and randomized trials comparing pressure garments, silicone, and combined treatment.

Selected references. Table 8 lists the top-10 papers in the selected source set. These top selected papers already cover several citation functions needed by a biomedical review: randomized clinical evidence, prior evidence synthesis, mechanistic evidence, and practical guideline context. This is important because a pressure-therapy meta-analysis needs not only intervention trials, but also scar-scale definitions, treatment guidelines, and

Stage	Hits	Precision	Recall
Direct title-only BM25 @10	3	0.300	0.061
Berry-LR prior ranking @10	7	0.700	0.143
Direct title-only BM25 @50	5	0.100	0.102
Berry-LR prior ranking @50	18	0.360	0.367
Direct title-only BM25 @1000	17	0.017	0.347
Berry-LR prior ranking @1000	46	0.046	0.939
Prior top-50 before selection	18	0.360	0.367
Sequential selected top-50	29	0.580	0.592

Table 7: Biomedical SurveyGen case-study result under the title-only protocol. Direct title-only BM25 uses only the target survey title. Berry-LR starts from the same title-only external input and uses internally generated query-planning text only as auxiliary model state. The final row shows the compact source set after trajectory-aware sequential reference selection.

mechanistic studies explaining why compression may alter scar tissue.

Takeaway. This case suggests that Berry-LR’s berrypicking evidence is not limited to computer-science topics. For a biomedical meta-analysis, direct title-only BM25 retrieval finds only 5 gold papers in the top 50, whereas the trajectory-derived prior ranking finds 18. The sequential selector then improves the compact 50-paper set to 29 gold references, raising F1 from 0.364 to 0.586. Qualitatively, the selected set is also coherent: it includes core randomized trials, pressure-garment and silicone comparisons, prior scar-management syntheses, preclinical compression evidence, and scar-assessment instruments. The case therefore serves as a useful appendix example of Berry-LR transferring to a biomedical literature-review setting.

D Case Study: Adaptive Berrypicking Retrieval Trace

This appendix provides a process-level case study for the SurveyBench topic *3D Object Detection for Autonomous Driving: A Comprehensive Survey*. The purpose is to make the berrypicking retrieval trace reproducible and inspectable: for each survey move, we report the generated query state, the anchors selected for subsequent expansion, the local retrieved candidates, and the evidence signals used by the trajectory-derived prior ranking. The run discovers 6461 unique candidate papers and evaluates the final top-1500 list, recovering 205/315 references from the human-written survey bibliography and 202/441 references from the benchmark

reference set. Unlike a one-shot retriever, Berry-LR keeps a separate retrieval state for each survey move and updates that state through anchor feedback.

D.1 Move-Round Retrieval Trace

Table 9 reports the complete move-round trace for the case. Each row corresponds to one survey move in one berrypicking round. We show the query intent used in that move-round, representative anchors selected for expansion, representative local retrieved candidates, and compact retrieval-control signals. The signal column reports anchor alignment, query-support ratio, and the number of local candidates that overlap with the human and benchmark reference sets.

D.2 Anchor-Conditioned Query Evolution

Table 10 summarizes how anchor feedback changes the query state across rounds. The table should be read from left to right: a representative anchor selected in one round is used, together with the survey specification and previous queries, to generate the next round’s move-conditioned query state. This makes the expansion more specific while preserving the intended survey function of the move. For example, the methods move shifts from broad model and multimodal retrieval toward LiDAR/fusion/point-cloud backbones and then toward PointFusion/VPFNet-style fusion methods.

D.3 Stopping and Continuation Criteria

The 3D object detection case was run with a three-round expansion budget, but Berry-LR is not defined as a fixed-depth procedure. After each move-round, the system first checks whether the eligible anchor pool contains papers whose anchor scores exceed τ_{anchor} . It then applies drift control to the temporary top- L anchors from that pool. If the eligible pool is empty or the drift-control indicator is triggered, the corresponding move has $\mathcal{H}_m^{(r)} = \emptyset$ and is no longer expanded.

D.4 Retrieval Evidence Accumulation

Table 11 shows how the final trajectory-derived prior ranking is supported by retrieval-trajectory evidence. High-ranked papers are not only similar to the initial topic: they are rediscovered by many query paths, cover multiple survey moves and rounds, appear repeatedly among local candidates, and often serve as anchors for later expansion. These quantities are the observable com-

Rank	Selected paper	Role	Gold	Direct@50
1	Pressure Garment Therapy Alone and in Combination with Silicone for the Prevention of Hypertrophic Scarring: Randomized Controlled Trial with Intraindividual Comparison	Classic method	Yes	Yes
2	Burn Scar Biomechanics after Pressure Garment Therapy	Representative application	Yes	Yes
3	The Use of Pressure and Silicone in Hypertrophic Scar Management in Burns Patients: A Pilot Randomized Controlled Trial	Representative application	Yes	Yes
4	A Randomized Clinical Trial to Study the Effect of Silicone Gel Dressing and Pressure Therapy on Posttraumatic Hypertrophic Scars	Representative application	Yes	Yes
5	A Translational Animal Model for Scar Compression Therapy Using an Automated Pressure Delivery System	Foundational concept	Yes	No
6	Pressure therapy for the control of hypertrophic scar formation after burn injury. A history and review	Prior survey / taxonomy	Yes	No
7	Development of a Best Evidence Statement for the Use of Pressure Therapy for Management of Hypertrophic Scarring	Prior survey / taxonomy	Yes	Yes
8	Updated scar management practical guidelines: non-invasive and invasive measures	Prior survey / taxonomy	Yes	No
9	In vitro mechanical compression induces apoptosis and regulates cytokines release in hypertrophic scars	Foundational concept	Yes	No
10	Prospective, randomized study of the efficacy of pressure garment therapy in patients with burns	Classic method	Yes	No

Table 8: Top-10 selected references in the biomedical case. Direct@50 indicates whether the same paper appears in the top 50 of direct title-only BM25 retrieval. Berry-LR keeps several gold references that direct title-only retrieval ranks outside its top-50 list, including prior synthesis, mechanistic, and randomized-trial evidence.

ponents of the retrieval evidence graph used by Berry-LR.

Summary. This case illustrates Berry-LR as an adaptive retrieval process rather than a single ranked retrieval call. Survey moves create separate retrieval states, anchors transfer information from one round to the next, stopping criteria limit uncontrolled drift, and the final ranking rewards papers supported by repeated and diverse retrieval evidence.

Round	Move	Query intent	Selected anchors	Representative candidates
1	Background	foundational surveys, perception basics, LiDAR/camera roots	3D Object Detection from Images for... [H+B]; 3D Object Detection for Autonomous... [H+B]	3D Object Detection from Images for... [H+B]; Survey and Systematization of 3D Object D...
1	Purpose	scope, task framing, taxonomy, application boundaries	3D Object Detection from Images for... [H+B]; Survey and Systematization of 3D Object D...	3D Object Detection from Images for... [H+B]; Survey and Systematization of 3D Object D...
1	Methods	models, architectures, LiDAR/camera/multimodal routes	An Overview Of 3D Object Detection [B]; AA3DNet Attention Augmented Real Time 3D...	Survey and Systematization of 3D Object D...; Multi-Modal 3D Object Detection in... [H+B]
1	Results	datasets, benchmarks, metrics, empirical comparisons	Survey and Systematization of 3D Object D...; 3D Object Detection from Images for... [H+B]	Survey and Systematization of 3D Object D...; 3D Object Detection from Images for... [H+B]
1	Limitations	failure modes, robustness, reliability, open problems	3D Object Detection from Images for... [H+B]; 3D Object Detection for Autonomous... [H+B]	3D Object Detection from Images for... [H+B]; 3D Object Detection for Autonomous... [H+B]
1	Conclusion	applications, trends, historical works, future directions	3D Object Detection from Images for... [H+B]; Survey and Systematization of 3D Object D...	3D Object Detection from Images for... [H+B]; Survey and Systematization of 3D Object D...
2	Background	follow MV3D/GS3D/Stereo R-CNN toward core methods	Multi-View 3D Object Detection Netw... [H+B]; GS3D An Efficient 3D Object Detection... [H+B]	3D Object Detection from Images for... [H+B]; 3D Object Detection for Autonomous... [H+B]
2	Purpose	use method anchors to refine scope, taxonomy, safety goals	GS3D An Efficient 3D Object Detection... [H+B]; Stereo R-CNN based 3D Object Detection... [H+B]	3D Object Detection from Images for... [H+B]; 3D Object Detection for Autonomous... [H+B]
2	Methods	expand to LiDAR, fusion, point-cloud backbones	RUHSNet 3D Object Detection Using Lidar D...; Cross-Modality 3D Object Detection [H+B]	RoIFusion 3D Object Detection from Li... [B]; Survey and Systematization of 3D Object D...
2	Results	query metrics and benchmarks around detector anchors	Complexer-YOLO Real-Time 3D Object... [H+B]; On the Robustness of 3D Object Detectors	Survey and Systematization of 3D Object D...; Complexer-YOLO Real-Time 3D Object... [H+B]
2	Limitations	follow localization-error and detector anchors to failures	Delving into Localization Errors fo... [H+B]; 3D Object Proposals using Stereo Im... [H+B]	Survey and Systematization of 3D Object D...; On the Robustness of 3D Object Detectors
2	Conclusion	connect overview/anchoring works to historical progress	An Overview Of 3D Object Detection [B]; Class-specific Anchoring Proposal for 3D...	Survey and Systematization of 3D Object D...; 3D Object Detection from Images for... [H+B]
3	Background	follow fusion anchors to RGB/point-cloud perception roots	Cross-Modality 3D Object Detection [H+B]; An Overview Of 3D Object Detection [B]	Survey and Systematization of 3D Object D...; Cross-Modality 3D Object Detection [H+B]
3	Purpose	expand to cooperative sensors and transportation context	High-level camera-LiDAR fusion for 3D obj...; Cooperative Perception for 3D Object Dete...	Survey and Systematization of 3D Object D...; Multi-Modal 3D Object Detection in... [H+B]
3	Methods	follow PointFusion/VPFNet to fusion and real-time backbones	PointFusion Deep Sensor Fusion for... [H+B]; 3D Object Detection Using Scale Invar... [B]	DeepFusion A Robust and Modular 3D Object...; Survey and Systematization of 3D Object D...
3	Results	open-set, voxel efficiency, datasets, robustness evidence	3D Object Detection Using Scale Invar... [B]; Open-set 3D Object Detection [H]	Survey and Systematization of 3D Object D...; On the Robustness of 3D Object Detectors
3	Limitations	cross-domain generalization and challenging environments	Train in Germany, Test in The USA Mak... [H]; RTS3D Real-time Stereo 3D Detection from...	Survey and Systematization of 3D Object D...; 3D Object Detection from Images for... [H+B]
3	Conclusion	long-term object-detection surveys and multi-sensor trends	Object Detection in 20 Years A Survey [B]; Multi-Task Multi-Sensor Fusion for... [H+B]	Survey and Systematization of 3D Object D...; Class-specific Anchoring Proposal for 3D...

Table 9: Move-round retrieval trace for the 3D object detection case. Each row reports the move-conditioned query intent, representative anchors selected for subsequent expansion, and representative candidates from the local fused retrieval results. Bracketed H/B markers after paper titles indicate whether that paper appears in the human-written reference set (H), the benchmark reference set (B), or both.

Move	R1 anchor	R2 query state	R2 anchor	R3 query state	R3 anchor
Background	3D Object Detection from Images for A... [H+B]	follow MV3D/GS3D/Stereo R-CNN toward core methods	Multi-View 3D Object Detection Networ... [H+B]	follow fusion anchors to RGB/point-cloud perception roots	Cross-Modality 3D Object Detection [H+B]
Purpose	3D Object Detection from Images for A... [H+B]	use method anchors to refine scope, taxonomy, safety goals	GS3D An Efficient 3D Object Detection... [H+B]	expand to cooperative sensors and transportation context	High-level camera-LiDAR fusion for 3D objec...
Methods	An Overview Of 3D Object Detection [B]	expand to LiDAR, fusion, point-cloud backbones	RUHSNet 3D Object Detection Using Lidar Dat...	follow PointFusion/VPFNet to fusion and real-time backbones	PointFusion Deep Sensor Fusion for 3D... [H+B]
Results	Survey and Systematization of 3D Object Det...	query metrics and benchmarks around detector anchors	Complexer-YOLO Real-Time 3D Object De... [H+B]	open-set, voxel efficiency, datasets, robustness evidence	3D Object Detection Using Scale Invaria... [B]
Limitations	3D Object Detection from Images for A... [H+B]	follow localization-error and detector anchors to failures	Delving into Localization Errors for... [H+B]	cross-domain generalization and challenging environments	Train in Germany, Test in The USA Makin... [H]
Conclusion	3D Object Detection from Images for A... [H+B]	connect overview/anchoring works to historical progress	An Overview Of 3D Object Detection [B]	long-term object-detection surveys and multi-sensor trends	Object Detection in 20 Years A Survey [B]

Table 10: Anchor-conditioned query evolution across rounds. Each row shows how a representative anchor from the previous round helps define the next query state for the same survey move. This table illustrates that later retrieval is generated from the survey specification, the move, previous queries, and selected anchors rather than from the title alone.

Rank	Paper	Gold	Paths	Moves	Rounds	Top	Anch.	Base
1	Survey and Systematization of 3D Object Detection M...	-	72	6	3	18	5	5
2	Multi-Modal 3D Object Detection in Autonomous Drivi...	H+B	72	6	3	17	5	3
3	3D Object Detection from Images for Autonomous Driv...	H+B	72	6	3	16	5	2
4	3D Object Detection for Autonomous Driving A Sur-vey	H+B	72	6	3	14	5	1
5	Multi-View 3D Object Detection Network for Au-tonomo...	H+B	72	6	3	15	4	11
6	3D Object Proposals using Stereo Imagery for Ac-cura...	H+B	71	6	3	15	4	6
7	GS3D An Efficient 3D Object Detection Framework for...	H+B	72	6	3	16	5	10
8	Stereo R-CNN based 3D Object Detection for Au-tonomo...	H+B	70	6	3	10	4	9
9	Accurate Monocular Object Detection via Color-Embed...	H+B	71	6	3	11	2	7
10	RoIFusion 3D Object Detection from LiDAR and Vi-sion	B	72	6	3	10	4	37
11	Pseudo-LiDAR from Visual Depth Estimation Bridg-ing...	H+B	70	6	3	12	2	13
12	Stereo CenterNet based 3D Object Detection for Auto...	B	71	6	3	7	1	12

Table 11: Retrieval evidence accumulated by representative top-ranked papers. **Paths** count query-level rediscoveries; **Moves** and **Rounds** measure trajectory coverage; **Top** counts appearances in move-round local candidate lists; **Anch.** counts anchor selections; **Base** is the rank from direct topic-based retrieval when available.