DOCAGENTIC-R: LEARNING UTILITY-AWARE RETRIEVAL FOR DOCUMENT DEEP RESEARCH

Anonymous authors
Paper under double-blind review

ABSTRACT

Training retrievers with downstream utility has improved agentic search, but existing frameworks assume homogeneous text passages. We study the same idea when candidates are text-carried document units that differ in modality and granularity. DocAgentic-Rtrains a BGE-M3 retriever with local relevance (LR), a prefix-freeze suffix-replan answer-contribution score (GAC), optional chain and localization labels, and multi-positive contrastive learning. On document-disjoint BlueprintSynth-Eval (2,430 questions; single runs), the full stack raises full-chain@20 from 0.588 to 0.674 and MRR from 0.713 to 0.800, above frozen Qwen3-Embedding-4B at 0.605. LR-only training falls below the backbone; GACis the signal that restores it. On the same pool, simpler positive rules reach 0.678–0.681, and text-only positives fall to 0.657. External transfer is about one point: MMDocIR Page R@5 moves from 0.705 to 0.710 and Layout R@5 stays below the backbone; the best LongDocURL setting gains 1.7 R@1 points and 1.0 LLM-judged point. Figure-unit coverage does not rise. Utility supervision helps synthetic multi-hop chain recovery. It does not, by itself, beat simpler utility labels or a native visual encoder.

1 INTRODUCTION

Deep research has become the dominant interface for knowledge work: instead of issuing a single query and reading a ranked list, an agent iteratively decomposes a question, issues sub-queries, reads the returned evidence, and decides what to ask next until it can produce a cited report. In this loop, retrieval is not a preprocessing step but the action space of the agent—every reasoning step is conditioned on what the retriever just returned, and a retrieval error compounds across turns. It is therefore surprising that the retriever itself is almost never the object of training: the system is taught to *cope with* a frozen retriever rather than to *improve* it.

Training retrievers has nonetheless progressed through three generations of supervision. The first generation optimizes *static relevance*: contrastive encoders such as DPR, E5, and BGE (Karpukhin et al., 2020; Wang et al., 2022; Xiao et al., 2023) learn to match a query to passages that are topically about it, with no notion of what the text will be used for. The second generation optimizes *single-turn utility*: REPLUG and LLM-Embedder (Shi et al., 2024; Zhang et al., 2024) use frozen-LM feedback such as generation likelihood or output rank to supervise retrieval. The third generation, most recently Agentic-R (Liu et al., 2026), optimizes *multi-turn utility*: it combines local relevance (LR) with the correctness of an answer generated after forcing a candidate passage into the current trajectory. Our GACscores a candidate by the drop in checklist satisfaction after that unit is removed and the suffix is replanned. This is a design difference from forcing the candidate in and reading out binary correctness; Section 5.2 does not show that the replan score beats simpler positive rules under a matched pool. Across all three generations, the retrieval unit in prior work is a homogeneous plain-text passage.

This assumption breaks precisely where agentic search is now heading. Document systems such as Doc-Researcher, MDocAgent, M3DocRAG, and ViDoRAG (Dong et al., 2026; Han et al., 2025; Cho et al., 2024; Wang et al., 2025) operate over PDFs and retrieve text or page-level visual evidence. Doc-Researcher's ablations show complementary gains from hybrid retrieval, adaptive planning, and iterative search; they do not isolate retriever training as the dominant bottleneck. Existing systems mainly keep first-stage retrievers off the shelf while improving surrounding components:

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

DocDancer (Zhang et al., 2026) uses supervised fine-tuning to learn an agent's Search/Read behavior, HiEvi-RAG (Xiong et al., 2026) trains an evidence verifier, and A-RAG (Du et al., 2026) exposes hierarchical retrieval interfaces. This leaves open how to train the first-stage retriever on downstream trajectory utility.

We argue this is not an oversight but a consequence of four structural differences that make document retrieval a distinct learning problem. (i) *Heterogeneous units*: candidates differ in modality (text, table, figure) and granularity (chunk, page, summary), so a score that is meaningful for one unit is not automatically comparable to another. (ii) *Chain-complementary value*: a unit's utility is often not intrinsic but relational—a table of Q3 revenue is only useful once the agent has the Q2 baseline, so relevance must be conditioned on what is still missing. (iii) *Grounding responsibility*: the output is a report with page/bbox citations, so retrieving the right content on the wrong page is a hard error, not a soft miss. (iv) *Parsing noise*: the candidate pool is produced by OCR and layout models, so the retriever must be robust to units that are degraded versions of each other.

We ask a single question: *can a retriever be trained on the downstream utility of heterogeneous document evidence, and where does that training stop helping?* DocAgentic-Rsupports two claims:

- **Chain recovery.** On a document-disjoint synthetic protocol, utility training raises full-chain recovery over frozen BGE-M3 and over frozen Qwen3-Embedding-4B. Relative to a text-only positive set that uses the same GACscore, the gain tracks keeping non-text units in the pool (section 5.2).
- **Boundary.** LR-only training falls below the backbone; GACremoves that drop. Chain labels, localization, typed hard negatives, and a second training round do not add a reliable gain. External page retrieval and frozen-agent answering move by about one point (section 5).

2 Related Work

Retriever training objectives. A first generation trains contrastive encoders on *static relevance*: DPR and its descendants, including E5 and the BGE family (Karpukhin et al., 2020; Wang et al., 2022; Xiao et al., 2023; Chen et al., 2024), learn to match a query to passages using labeled or weakly supervised pairs. A second generation optimizes *single-turn utility*: REPLUG and LLM-Embedder (Shi et al., 2024; Zhang et al., 2024) distill frozen-LM feedback such as generation likelihood or output rank into a retriever. The third generation, exemplified by Agentic-R (Liu et al., 2026), moves to *multi-turn utility*: it evaluates final-answer correctness after forcing each candidate passage into the current trajectory and combines that binary signal with local relevance. Our GACinstead removes the candidate under a prefix freeze and replans the suffix; the score is the drop in checklist satisfaction. We do not treat that formula as a result that outperforms Agentic-R. These methods have been developed and evaluated primarily with homogeneous plain-text passages, rather than units that differ in modality, granularity, or grounding responsibility.

Document deep research and document agents. A separate line of work builds *systems* that operate on long documents while using off-the-shelf first-stage retrievers. Doc-Researcher (Dong et al., 2026) unifies multimodal parsing, adaptive planning, and iterative search; its ablations show gains from all three rather than isolating retriever training as the dominant bottleneck. MDocAgent (Han et al., 2025), M3DocRAG (Cho et al., 2024), and ViDoRAG (Wang et al., 2025) orchestrate text or visual retrievers with multi-agent or multi-page reasoning. Where these systems train agentic components, DocDancer (Zhang et al., 2026) uses supervised fine-tuning over synthesized Search/Read trajectories, HiEvi-RAG (Xiong et al., 2026) trains a multi-page evidence verifier with GRPO, and A-RAG (Du et al., 2026) exposes hierarchical retrieval interfaces without training the underlying retriever. This line therefore does not directly optimize first-stage document retrieval from downstream multi-turn utility.

Visual document retrieval. A third line trains retrievers that see document layout, but with query–document relevance supervision. ColPali and its successors (Faysse et al., 2025) adapt ColBERT-style late interaction to page screenshots; jina-embeddings-v4 (Günther et al., 2025) produces universal multimodal embeddings. They are strong frozen visual baselines in our evaluation, whereas

Table 1: Positioning of DocAgentic-Ragainst the closest lines of work.

	Trained part	Retrieval unit	Supervision	Evidence chain
Agentic-R	retriever	homogeneous text chunk	LR+GAC	none
DocDancer	agent policy	Search/Read trajectories	supervised fine-tuning	none
HiEvi-RAG	evidence verifier	page image	grounding GRPO	verification only
A-RAG	none (interfaces)	keyword/sentence/chunk	—	none
DocAgentic-R(ours)	retriever	text-carried, multi-granular	LR+GAC+ECC+LOC	in training signal

our trainable text-carrier retriever is initialized from BGE-M3. Their supervision contains no explicit evidence-chain contribution or downstream trajectory-utility term.

3 Problem Formulation

We formalize document deep research as multi-turn retrieval over heterogeneous evidence units, then define the learning problem.

Heterogeneous retrieval units. A parsed document corpus is not a set of passages but a pool of *evidence units*. Each unit is a tuple

$$e = (c, m, g, d, p, b), \tag{1}$$

where c is the content representation, $m \in \{\text{text}, \text{table}, \text{figure}, \text{equation}\}$ is the layout modality, $g \in \{\text{chunk}, \text{page}, \text{summary}\}$ is the granularity, and (d, p, b) are the document id, page number, and bounding box. The content c is itself multi-faceted: a text chunk is its token sequence, a table or figure is represented by a VLM-generated caption (this submission does not add page renders or train ColQwen/multi-vector models), and a page summary is an abstractive condensation. Crucially, the same underlying fact can be instantiated as several units whose utilities differ—a Q3 revenue figure may exist as a table cell, a chart, and a sentence in the MD&A—so the choice of representation is not a preprocessing decision but part of what must be learned.

Multi-turn agent trajectories. An agent interacts with retriever $\mathcal{R}$ over a corpus $\mathcal{E}$ of such units. Given a question Q, it issues a short sequence of sub-queries—typically $T \approx 3$–5 turns in the trajectories used for supervision—and consumes the retrieved evidence before answering:

$$\mathcal{T} = [(q_1, \mathcal{C}_1), \ldots, (q_T, \mathcal{C}_T)], \qquad \hat{a} = \text{Agent}(Q, \mathcal{T}), \tag{2}$$

where q_i is the agent-issued sub-query at turn i and $\mathcal{C}_i = \mathcal{R}(x_i, \mathcal{E}) \subset \mathcal{E}$ is the retrieved set. Unlike in text QA, $\hat{a}$ is not just a string but a report with an explicit localization set $\hat{L} = \{(\hat{p}_k, \hat{b}_k)\}$ of cited pages and bounding boxes. The retriever's output is therefore consumed in two ways: as context for the next reasoning step, and as the citation basis of the final artifact.

The default query-side encoding used for training and for all main tables is the original question concatenated with the current sub-query:

$$x_i = Q\,[\text{SEP}]\,q_i. \tag{3}$$

An optional evidence-state summary S_{i-1} ("which link is still missing") can be appended as $x_i = Q\,[\text{SEP}]\,q_i\,[\text{SEP}]\,S_{i-1}$. Table 14 shows that neither a raw dump of prior evidence nor a gold-free gap summary improves full-chain recovery once train and evaluation forms match, so S_{i-1} is defined but left off by default.

Learning objective and replan intervention. We learn a scoring function $s_\theta(x_i, e)$ whose goal is not to approximate a static relevance label but to maximize the expected downstream utility of $\mathcal{C}_i$, under the constraint that scores remain comparable across unit types. Evidence chains are treated as *sets*, matching the available annotations.

The global answer-contribution signal used for that objective is computed on the same multi-turn trajectories. After intervening on candidate e at turn i, we *freeze* $q_{\le i}$ and the retrieved prefix $\mathcal{C}_{<i}$, then let the same frozen agent regenerate later sub-queries $q_{>i}$ and the final answer. Units that cannot complete a replan fall back to fixed-set leave-one-out; the scoring details are in section 4.3.

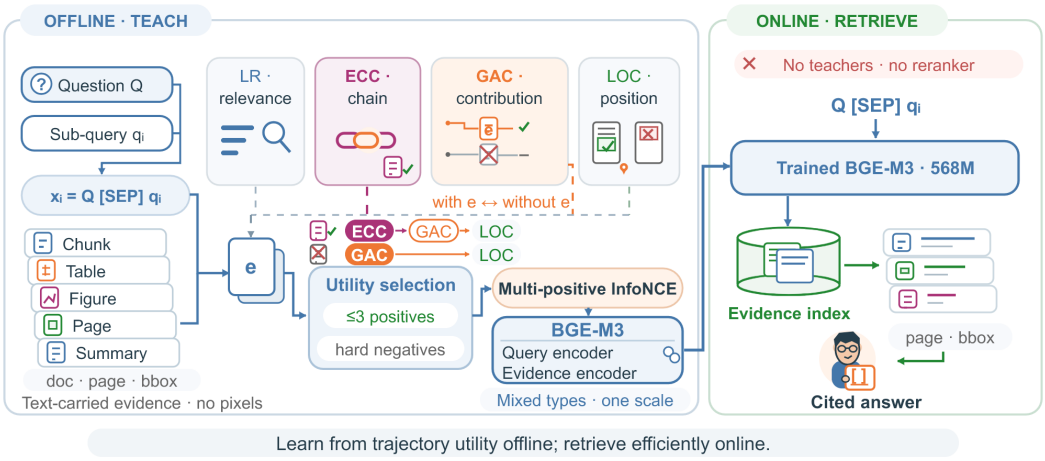

Figure 1: Overview of DocAgentic-R. Offline LR, counterfactual GAC, chain-based ECC, and localization-based LOC select multi-positive training units for a shared BGE-M3 dual encoder. At deployment, only the trained encoder and a precomputed mixed-evidence index are used; no teacher or online reranker is required.

4 DocAgentic-R

We describe the encoder, the four supervision signals, the contrastive objective, and how training data is built.

4.1 Retriever model form

The trainable object must satisfy one hard constraint: competition among evidence units has to happen at recall time, over a precomputed index, without an online reranker. This rules out two natural alternatives. A cross-encoder reranker sees query and candidate jointly and cannot serve the first-stage recall of 10^5–10^6 units; a generative VLM produces no indexable representation at all. We therefore train a dense dual-tower encoder in the 0.1–0.6B range, where the document side is encoded once and cached. All teachers—the listwise LRscorer, the trajectory agent for GAC—exist only at signal-construction time; at deployment only the small encoder is online. We initialize from BGE-M3 (Chen et al., 2024) (568M, 8192-token context), which comfortably fits the three-part input of Eq. 3; Appendix D gives the full rationale and a comparison against cross-encoder and multi-vector candidates.

4.2 Cross-unit comparability

All units are serialized with modality and granularity metadata and scored by one encoder. The denominator of Eq. 7 mixes unit types. Figures and tables enter as VLM captions, so this tests cross-granularity competition more than native visual retrieval. Extra type embeddings and quantile biases lower full-chain@20 from 0.674 to 0.659; text prefixes recover 0.678. Neither extra module is used in the main tables.

4.3 Four supervision signals

For each (sub-query q_i, candidate unit e) pair we compute four signals.

(1) LR– local relevance. A frozen text LLM gives listwise scores in $[0, 100]$ with layout/granularity labels injected into the prompt. Visual units are represented by their VLM-generated captions in v1; consequently, LRpartly measures caption quality, a limitation we isolate in section 6.

LRis a fault-tolerant *last-key* signal: it coarse-ranks the pool, while GAC/ECC/LOCdetermine positive membership.

(2) GAC– global answer correctness. The current implementation measures marginal contribution on multi-turn trajectories by freezing the prefix and replanning the suffix. After removing candidate e at turn i, we freeze $q_{\le i}$ and the retrieved prefix, then let the same frozen agent regenerate later sub-queries and answer. Let $\hat{a}$ be the original-trajectory answer and $\hat{a}_{-e}$ the answer after the intervened replan:

$$\mathrm{GAC}(e) = J\big(\hat{a}, a^*\big) - J\big(\hat{a}_{-e}, a^*\big), \tag{4}$$

where J is checklist satisfaction for long document answers. Positive values indicate useful evidence, values near zero indicate redundancy, and negative values indicate distractors. The rollout budget is tilted toward early turns. Units that cannot complete a replan fall back to fixed-set leave-one-out; units still without a valid record fall back to the earlier single-unit score.

(3) ECC– evidence-chain contribution. Given a gold evidence chain $C^* = \{e_1^*, \ldots, e_K^*\}$,

$$\mathrm{ECC}_i(e) = \max\Big(\mathbf{1}[e \in C^*],\ \lambda \cdot \Delta\mathrm{Cover}_i(e)\Big), \qquad \Delta\mathrm{Cover}_i(e) = \big|C^* \cap (\mathcal{C}_{\le i} \cup \{e\})\big| - \big|C^* \cap \mathcal{C}_{\le i}\big|. \tag{5}$$

Chain membership is non-revocable (annotation decides); complementarity is a bonus channel for unannotated units. Motivation: in multi-hop document questions, value is often complementary—a unit that completes the missing link beats one that is individually "correct".

(4) LOC– localization correctness (new).

$$\mathrm{LOC}(e) = \max_{(\hat{p},\hat{b}) \in \hat{L}} \Big[\alpha \cdot \mathbf{1}[\hat{p} = p_e] + (1-\alpha) \cdot \mathrm{IoU}(\hat{b}, b_e)\Big]. \tag{6}$$

LOCpenalizes "right content, wrong page", which is harmless in text QA but a hard error in document research reports with bbox citations. Its independent value concentrates in three document-specific decoupling cases: same-doc adjacent pages, cross-doc same-entity units, and VLM-caption attribution drift ($\alpha \in [0.5, 0.7]$ with bbox gold; α=1 for page-only gold).

Signal combination as a decision tree.

- *Samples with chain annotations (weak or cold-start):* ECCis the primary key; GACacts as a gate (chain units with GAC ≤ 0 are demoted to weak positives).
- *Samples without chains:* GACis the primary key with multi-judge voting (two LLM families + rule checks) to reduce variance.
- LOCalways enters as the final tie-breaker; LRranks the pool.

4.4 Multi-positive contrastive training

A single document sub-query often requires several complementary units (table value + figure trend), so we replace the one-positive objective with

$$\mathcal{L} = -\log \frac{\sum_{e^+ \in \mathcal{P}} \exp\big(s_\theta(x, e^+)/\tau\big)}{\sum_{e^+ \in \mathcal{P}} \exp\big(s_\theta(x, e^+)/\tau\big) + \sum_{e^- \in \mathcal{N}} \exp\big(s_\theta(x, e^-)/\tau\big)}, \tag{7}$$

For samples with chain annotations, $\mathcal{P}$ contains up to three chain units that pass the GACgate; for samples without chains, it contains up to three candidates selected by GACwith multi-judge voting. Thus ECCdefines candidate membership only when chain labels exist. In both cases $|\mathcal{P}| \le 3$ prevents gradient dilution, $|\mathcal{P}| + |\mathcal{N}| = 16$, and $\tau = 0.01$, with in-batch and cross-device negatives.

4.5 Negatives and granularity

The default mix uses same-document neighbors and off-chain distractors. Five typed near-misses were implemented (Table 15); leave-one-type-out on the present corpus does not isolate any type as necessary (Table 16). Chunk, page, and summary units share one pool. Hard-filtering granularity loses about 10 full-chain@20 points for both the full model and the frozen encoder of that run (Table 7), so the loss is from dropping page and summary units, not from utility training.

4.6 TRAINING DATA CONSTRUCTION

Training data is a three-way merge: ColdStart(MMDocIR Train$\cup$CMDR-Synth)$\cup$BlueprintSynth$\setminus$ Dedup(Eval). BlueprintSynth first samples evidence blueprints from public distributional priors (text/figure/table/equation $\approx$ $60/19/17/4$, target cross-document rate $\approx$ 40%, and chunk/page/summary $\approx 80/15/5$), matches concrete evidence chains, and then asks a frozen LLM to generate multi-hop questions *backward* from those chains. The chain is therefore an input to synthesis and an exact synthetic label, not a weak label inferred from citation logs. Questions answerable by the base model without retrieval are removed by a pass@10 filter. MMDocIR Eval, LongDocURL, and CMDR-Bench are fully held out; BlueprintSynth-Eval uses 1,200 documents disjoint from the training questions. The calibration subset (seed 20260728) is used only for model selection.

5 EXPERIMENTS

We test chain recovery on document-disjoint BlueprintSynth-Eval, transfer on MMDocIR, LongDocURL, and CMDR-Bench, and whether extra signals beat simpler positive rules. All numbers are single runs.

5.1 SETUP

MMDocIR Eval (Dong et al., 2025) scores page and layout retrieval. LongDocURL (Deng et al., 2025) scores evidence-page retrieval and frozen-agent answers. CMDR-Bench (Tanaka et al., 2026) scores held-out indirect page retrieval. BlueprintSynth-Eval is a 2,430-question, document-disjoint multi-hop protocol that shares the training generator and prior. Controlled baselines are frozen BGE-M3, Qwen3-Embedding-4B (Zhang et al., 2025), and page-image jina-embeddings-v4. Full metric sets and hyperparameters are in Appendix F and Appendix D.

5.2 CHAIN RETRIEVAL

BlueprintSynth-Eval uses 1,200 documents never seen in training, 2–3-hop chains, and 42% cross-document questions. It is an in-distribution mechanism test, not an external benchmark.

Table 2: BlueprintSynth-Eval chain retrieval (2,430 questions). Full-chain@20 is the headline metric.

Configuration	Unit R@10	Chain R@20	Full-chain@20	nDCG@10	MRR
BGE-M3 (base)	0.911	0.760	0.588	0.605	0.713
LRonly	0.955	0.801	0.629	0.627	0.738
LR+GAC	0.975	0.826	0.664	0.665	0.781
LR+GAC+ECC	0.978	0.832	0.672	**0.682**	**0.802**
Full (+LOC)	**0.979**	**0.833**	**0.674**	0.681	0.800

Relative to frozen BGE-M3, the full stack gains 8.6 full-chain@20 points and 8.7 MRR points. Three-hop questions rise from 0.335 to 0.385; cross-document questions rise from 0.516 to 0.616. Adding GACafter LR-only raises MRR by 4.3 points. Frozen Qwen3-Embedding-4B reaches 0.605 under the same static-Q protocol (2-hop 0.693, 3-hop 0.347): above the backbone, below the full stack.

5.3 EXTERNAL TRANSFER

Table 3 compresses the three held-out page protocols. The best utility row is the full stack on MMDocIR and CMDR, and LR+GAC+ECCon LongDocURL.

Table 3: Transfer on held-out page protocols. MMDocIR uses the same OCR as training. LongDocURL answers use a frozen Qwen3.6-27B on the top-5 pages. Full metric sets are Tables 11–13.

Model	MMDocIR P@5	Layout R@5	LongDoc R@1	LongDoc judge	CMDR R@1
BGE-M3 (base)	0.705	**0.580**	0.477	0.545	0.161
LRonly	0.685	0.466	0.452	0.534	0.149
Best utility	0.710	0.571	**0.494**	**0.555**	**0.178**
Qwen3-Emb-4B	0.734	0.432	–	–	–
Jina-v4 (page image)	**0.848**	–	–	–	–

Gains stay near one point. Layout R@5 stays below the backbone. Jina-v4's page-image lead is the visual-encoder gap: the trained retriever never encodes page pixels, so it cannot close that comparison. The same ranking appears inside MMDocIR (Table 4).

Table 4: Signal ranking on MMDocIR Eval (fair OCR, 1,658 questions).

Configuration	Page R@5	nDCG@10	Layout R@5
BGE-M3 (base)	0.705	0.585	**0.580**
LRonly	0.685	0.558	0.466
LR+ GAC	0.704	0.580	0.542
LR+ GAC+ ECC	0.708	**0.590**	0.565
Full (+LOC)	**0.710**	0.589	0.571

LR-only drops Layout R@5 by 11.3 points, so the collapse is not a page-only artifact. Replan GACreturns page metrics to the backbone and recovers most of the layout drop; adding ECCand LOCthen moves Page R@5 by half a point. The LongDocURL end-to-end ranking and the figure-coverage drop are in section 6.

5.4 What extra signals add

On the same mixed pool, simpler positive rules match or exceed the four-signal stack (Table 5).

Table 5: Same-pool positive rules on BlueprintSynth-Eval. Cited / leave-one-out / soft×chain use the mixed pool. Single runs.

Supervision	Full-chain@20
Full stack	0.674
Cited units	0.678
Leave-one-out GAC > 0	0.679
Soft LR $\times$ chain	0.681
Text-only + replan GAC	0.657
Coverage sweep, ρ=0 (pure GAC)	0.609
Coverage sweep, ρ=1 (pure ECC)	0.600

Text-only positives lose 1.7 points because 295 of 850 groups have no text positive and are dropped. The ρ rows come from a separate coverage sweep and must not be subtracted from the 0.674 full-stack row. Table 6 makes the same point at finer grain.

Table 6: Chain-label coverage sweep on BlueprintSynth-Eval. "Relative rollout" is the implied fraction of questions requiring GAC.

ρ	0.0	0.2	0.4	0.6	0.8	1.0
Relative rollout	1.0	0.8	0.6	0.4	0.2	0.0
nDCG@10	**0.639**	0.635	0.629	0.628	0.631	0.623
Full-chain@20	**0.609**	0.600	0.602	0.600	0.603	0.600
MRR	**0.761**	0.757	0.745	0.746	0.750	0.740

Pure GAC(ρ=0) is strongest, and replacing every chain label for a rollout (ρ=1) costs quality. Hard-filtering granularity is a pool decision, not a utility-training effect. Table 7 uses one encoding pass.

Table 7: Granularity-pool ablation on BlueprintSynth-Eval static Q (2,430 questions).

Model	**Pool**	**Chain R@20**	**Full@20**	**nDCG@10**	**MRR**
BGE-M3	Compete	0.827	0.665	0.673	0.790
BGE-M3	Planner	0.762	0.565	0.640	0.773
BGE-M3	chunk	0.764	0.565	0.642	0.775
BGE-M3	oracle	0.838	0.679	0.694	0.811
Full	Compete	0.833	0.674	0.681	0.801
Full	Planner	0.764	0.568	0.648	0.780
Full	chunk	0.766	0.569	0.650	0.782
Full	oracle	**0.840**	**0.685**	**0.699**	**0.816**

Planner and chunk-only drop about 10 full-chain@20 points for both the full model and frozen BGE-M3. The frozen compete row is 0.665, not the main-table backbone 0.588. Appending an evidence-state summary S_{i-1} is likewise not the default (Table 14 in the appendix). Typed negatives stay in Appendix F.

6 Discussion and Conclusion

On the full LongDocURL answer set a frozen 27B agent repeats the one-point transfer: LR-only drops 1.1 judge points, LR+GAC+ECCis best at +1.0, and exact match stays near 0.20 (Table 8).

Table 8: LongDocURL end-to-end on 2,317 questions. Frozen Qwen3.6-27B reads the same top-5 pages from each retriever.

Retriever	**LLM-judge**	**EM**	**F1**
BGE-M3 (base)	0.545	0.204	0.350
LRonly	0.534	0.205	0.347
LR+GAC	0.545	0.205	0.352
LR+GAC+ECC	**0.555**	**0.207**	**0.355**
Full (+LOC)	0.552	0.204	0.352

Chain recovery also does not raise figure coverage. Table 9 restricts each row to questions whose gold chain contains that modality: table and equation gold hits rise, figure gold hit falls from 0.773 to 0.764.

Table 9: Modality coverage on BlueprintSynth-Eval (static Q, top-20).

Modality	n	base gold	full gold	base any	full any
text	2127	0.910	**0.912**	0.995	0.995
table	576	0.816	**0.840**	0.844	**0.865**
figure	653	**0.773**	0.764	**0.824**	0.819
equation	179	0.358	**0.391**	0.564	**0.626**
all needed mods	2430	0.794	**0.798**	0.886	**0.893**

Figure 2 restates the same split: a large full-chain gain on the document-disjoint synthetic protocol, and about one point on external page and end-to-end metrics.

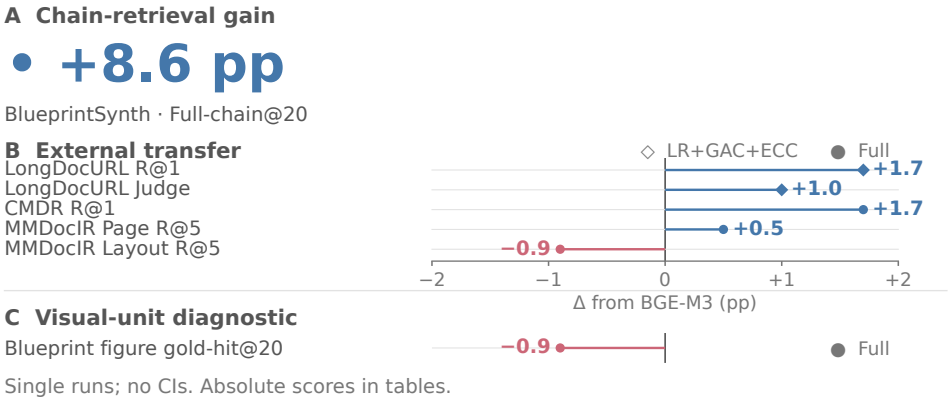

Figure 2: Boundary of the observed utility-supervision gains. Panel A: Full-chain@20 on BlueprintSynth-Eval. Panel B: changes from frozen BGE-M3 on external benchmarks. Panel C: in-distribution figure-unit diagnostic. Single-run descriptive differences, no confidence intervals.

Utility training raises full-chain recovery when the test asks for a synthetic multi-hop chain: 0.674 versus 0.588 for frozen BGE-M3 and 0.605 for frozen Qwen3-Embedding-4B. LR-only training falls below the backbone; GACremoves that drop. Mixed-pool rules that ignore the four-signal tree reach 0.678–0.681; text-only positives fall to 0.657. Typed negatives, a second training round, and an unseen subquery writer do not add a reliable gain (Appendix F).

Figures and tables enter as VLM captions, so retrieval partly scores caption quality. BlueprintSynth-Eval shares the generator and prior with training and uses only 2–3 hops. Units that cannot finish a replan fall back to leave-one-out. Every reported number is a single run.

A retriever can be trained on the downstream utility of text-carried document units. On this document-disjoint synthetic protocol, the gain over a text-only positive set comes from keeping non-text units in the pool. Extra signals beyond an answer-contribution score, and transfer beyond about one point on page protocols, are not established here.

REFERENCES

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*, 2024.

Jaemin Cho, Debanjan Mahata, Ozan Irsoy, Yujie He, and Mohit Bansal. M3docrag: Multi-modal retrieval is what you need for multi-page multi-document understanding. *arXiv preprint arXiv:2411.04952*, 2024.

Chao Deng, Jiale Yuan, Pi Bu, Peijie Wang, Zhong-Zhi Li, Jian Xu, Xiao-Hui Li, Yuan Gao, Jun Song, Bo Zheng, and Cheng-Lin Liu. LongDocURL: A comprehensive multimodal long document benchmark integrating understanding, reasoning, and locating. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pp. 1135–1159, 2025. doi: 10.18653/v1/2025.acl-long.57.

Kuicai Dong, Yujing Chang, Derrick Goh Xin Deik, Dexun Li, Ruiming Tang, and Yong Liu. MMDocIR: Benchmarking multimodal retrieval for long documents. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 30971–31005, 2025. doi: 10.18653/v1/2025.emnlp-main.1576.

Kuicai Dong, Shurui Huang, Fangda Ye, Wei Han, Zhi Zhang, Dexun Li, Wenjun Li, Qu Yang, Gang Wang, Yichao Wang, Chen Zhang, and Yong Liu. Doc-researcher: A unified system for multimodal document parsing and deep research. In *Proceedings of the ACM Web Conference 2026*, pp. 2349–2360, 2026. doi: 10.1145/3774904.3792599.

Mingxuan Du, Benfeng Xu, Chiwei Zhu, Shaohan Wang, Pengyu Wang, Xiaorui Wang, and Zhendong Mao. A-rag: Scaling agentic retrieval-augmented generation via hierarchical retrieval interfaces. *arXiv preprint arXiv:2602.03442*, 2026.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. Colpali: Efficient document retrieval with vision language models. In *International Conference on Learning Representations*, 2025.

Michael Günther, Saba Sturua, Mohammad Kalim Akram, Isabelle Mohr, Andrei Ungureanu, Sedigheh Eslami, Scott Martens, Bo Wang, Nan Wang, and Han Xiao. jina-embeddings-v4: Universal embeddings for multimodal multilingual retrieval. *arXiv preprint arXiv:2506.18902*, 2025.

Siwei Han, Peng Xia, Ruiyi Zhang, Tong Sun, Yun Li, Hongtu Zhu, and Huaxiu Yao. Mdocagent: A multi-modal multi-agent framework for document understanding. *arXiv preprint arXiv:2503.13964*, 2025.

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

Wenhan Liu, Xinyu Ma, Yutao Zhu, Yuchen Li, Daiting Shi, Dawei Yin, and Zhicheng Dou. Agentic-r: Learning to retrieve for agentic search. In *Findings of the Association for Computational Linguistics: ACL 2026*, 2026. doi: 10.18653/v1/2026.findings-acl.785.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. REPLUG: Retrieval-augmented black-box language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 8371–8384, 2024. doi: 10.18653/v1/2024.naacl-long.463.

Ryota Tanaka, Taku Hasegawa, and Kyosuke Nishida. CMDR: Contextual multimodal document retrieval. In *Computer Vision – ECCV 2026*, pp. 77–95, 2026. doi: 10.1007/978-3-032-37035-8_5.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*, 2022.

Qiuchen Wang, Ruixue Ding, Zehui Chen, Weiqi Wu, Shihang Wang, Pengjun Xie, and Feng Zhao. Vidorag: Visual document retrieval-augmented generation via dynamic iterative reasoning agents. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025. doi: 10.18653/v1/2025.emnlp-main.464.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. C-pack: Packaged resources to advance general chinese embedding. *arXiv preprint arXiv:2309.07597*, 2023.

Junyu Xiong, Yonghui Wang, Rongjian Gu, Chenyu Liu, Bing Yin, Wengang Zhou, and Houqiang Li. Hievi-rag: Hierarchical evidence-driven reasoning for long document understanding. *arXiv preprint arXiv:2607.04625*, 2026.

Peitian Zhang, Zheng Liu, Shitao Xiao, Zhicheng Dou, and Jian-Yun Nie. A multi-task embedder for retrieval augmented LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024. doi: 10.18653/v1/2024.acl-long.194.

Qintong Zhang, Xinjie Lv, Jialong Wu, Baixuan Li, Zhengwei Tao, Guochen Yan, Huanyao Zhang, Bin Wang, Jiahao Xu, Haitao Mi, and Wentao Zhang. Docdancer: Towards agentic document-grounded information seeking. *arXiv preprint arXiv:2601.05163*, 2026.

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

AI USE STATEMENT

Generative AI tools were used to organize the literature review, assist language editing and translation, generate synthetic questions from preselected evidence chains, caption document figures, score local relevance, and judge counterfactual/final answers. They were not treated as sources of factual claims: the authors checked citations, inspected generated data, and computed all reported metrics from saved model outputs. The authors take responsibility for the final text, claims, data pipeline, and artifacts.

ETHICS STATEMENT

The study uses public document corpora and benchmark datasets under their respective licenses; users of the released processing pipeline must separately verify the terms of each source. No human subjects or private user data are involved. Synthetic questions and VLM captions may inherit factual or demographic biases from their generators. LLM-based relevance and answer judges can also exhibit family-specific preferences; accordingly, we report their role explicitly and avoid treating judge scores as human ground truth. Document identifiers are removed from training when they match the evaluation denylist.

REPRODUCIBILITY STATEMENT

The submission package specifies the BGE-M3 initialization, data manifests, document-level evaluation denylist, blueprint-generation configuration, random seed 20260728, signal-construction scripts, checkpoint names, and machine-readable evaluation reports. All tables in the current draft are traceable to saved JSON reports. Prompts and additional implementation details are included in the appendix and will accompany the anonymized code release.

A DATASETS AND DECONTAMINATION DETAILS

Parsed corpus. The working corpus manifest contains 4,115 documents: 1,600 arXiv PDFs (`hybridfree/arxiv-papers`), 1,599 SEC filings (`PleIAs/SEC`), 462 EDGAR filings (`anonymous-md/EDGAR_FILINGS_DATASET`), and 454 DocLayNet pages (`docling-project/DocLayNet-v1.2`). Each document is parsed into the evidence units of

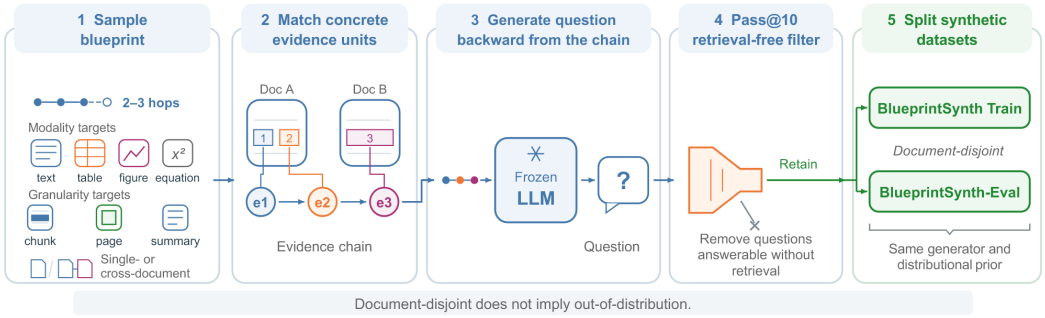

Figure 3: Backward construction of BlueprintSynth from matched evidence chains. We sample a structural blueprint, bind it to concrete evidence units, generate a question from the resulting chain, and discard questions answerable without retrieval under pass@10. The training and evaluation partitions are document-disjoint but share the same generator and distributional prior.

Eq. 1. Figure units store a VLM caption in addition to the image path; the current retriever never encodes page pixels.

Held-out evaluation. MMDocIR Eval, LongDocURL, and CMDR-Bench labels never enter training (section 4.6). An evaluation denylist of 727 keys, built from the MMDocIR Eval pages and LongDocURL documents present at construction time, is applied when assembling the training mixture. CMDR-Bench was not on disk when that denylist was written; its exclusion is enforced by the hold-out rule rather than by those 727 keys. BlueprintSynth-Eval uses 1,200 documents that are disjoint from the training questions.

Evaluation counts. The published MMDocIR paper reports 1,685 expert questions; the released annotation file used by our pipeline contains 1,658, all of which are evaluated. LongDocURL reports 2,325 questions. Our retrieval protocol evaluates the 2,323 entries with non-empty evidence-page labels, while the end-to-end protocol evaluates the 2,317 entries with non-empty gold answers. These filters are fixed before comparing retrievers.

Deduplication. Training questions are formed from ColdStart(MMDocIR Train $\cup$ CMDR-Synth) $\cup$ BlueprintSynth $\setminus$ Dedup(Eval). Exact evaluation documents and denylist keys are dropped before pair construction. We do not claim a measured near-duplicate rate beyond this exact-key filter.

B BLUEPRINT SYNTHESIS

BlueprintSynth first samples an evidence *blueprint*—a short specification of hop count, modalities, granularities, and whether the chain should cross documents—then matches concrete units in the parsed corpus and asks a frozen LLM to write a multi-hop question *backward* from that chain (section 4.6). The chain is therefore an input to synthesis and an exact synthetic label. figure 3 summarizes this construction and the document-disjoint split.

The current training dump contains 5,000 matched blueprints (all `match_ok`). Hop counts are 3,465 two-hop and 1,535 three-hop items. Cross-document blueprints number 2,076 (41.5%), close to the 40% sampling target. The intended modality and granularity priors are those stated in section 4.6: text/figure/table/equation $\approx 60/19/17/4$ and chunk/page/summary $\approx 80/15/5$. Questions answerable by the frozen base model without retrieval are removed by pass@10.

A typical record stores `qid`, the blueprint string (for example `table:chunk+text:chunk`), hop count, target document count, the matched unit ids, and a seed of 20260728. BlueprintSynth-Eval is a document-disjoint split of this generator, not a human-annotated multi-hop set; its 2–3 hop length is a limitation discussed in section 6.

C SIGNAL CONSTRUCTION PROMPTS

All scoring calls use a frozen Qwen-family text LLM with thinking disabled. Visual units enter as VLM captions, not page pixels. We do not run a second-family judge; a rule-based salient-token vote is recorded only as an analysis check.

Local relevance (LR). For each turn the judge sees the original question, the current sub-query, and the retrieved units (id, modality, page, and truncated text or caption). It assigns an integer in $[0, 100]$ per unit and must emit lines of the form `UNIT_ID: SCORE`. The score is a last-key coarse rank, not the positive-membership decision.

Global answer contribution (GAC). Checklist construction asks for 3–5 atomic facts that a complete answer must contain, each independently verifiable against the gold evidence, one item per line prefixed with `-`. The answer judge then sees the question, the candidate answer, the gold evidence, and that checklist, and must emit `ITEM` i`:` `yes` or `ITEM` i`:` `no`. J is the fraction of yes items. Abstaining rollouts are scored 0 without a judge call. Units that cannot complete prefix-freeze suffix-replan fall back as described in section 4.3.

What is not claimed. Weak-chain multi-trajectory voting and a second independent LLM judge are not part of the current pipeline. Evidence-chain contribution (ECC) and localization (LOC) are computed from annotations and box overlap, not from additional free-text prompts.

D IMPLEMENTATION DETAILS

The retriever is initialized from BGE-M3 (568M) and trained on one GPU with automatic mixed precision. We optimize the multi-positive InfoNCE objective of section 4.4 with $\tau = 0.01$ and 16 units per query (at most three positives plus hard negatives). Embeddings are L_2-normalized and the loss is evaluated in fp32 for numerical stability.

Captioning visual units. The parser initially stores a figure unit as an image-path placeholder, which carries no semantics for a text encoder. A frozen VLM generates a content caption for each figure, and the same replacement is applied to training and evaluation corpora. The current model does not encode page pixels directly.

Training mixture. BlueprintSynth provides the 16-unit utility-supervised groups. ColdStart pairs derived from the MMDocIR training split are mixed in to reduce the gap between cross-document synthesis and single-document evaluation. This mixture improves page retrieval but can reduce layout-block precision, as discussed in Appendix F.2.

E COST TRANSPARENCY

Table 10 lists wall-clock hours read from the `DONE` lines of the production logs (file `Data/eval/reports/b7_gpu_hours.json`). Language-model stages ran on a remote frozen Qwen3.6-27B endpoint, not on the local training GPUs. Retriever training ran on one local GPU per configuration. These numbers are not FLOP-normalized and do not include failed retries that left no `DONE` line.

The "Relative rollout" row of Table 6 is the fraction of *questions* that still need a GACrollout, $1 - \rho$. It is not a measured $1 - \rho$ drop in Table 10. Synthesis, LR, and training stay in the pipeline at every ρ. At $\rho = 0.6$ the implied rollout count is 40% of full GAC; the logged replan-GACstage is 4.34 hours, so that slice would be about 1.74 hours if the same per-question cost held. We do not treat that product as a new measurement.

Table 10: Logged wall-clock hours for the main v5 pipeline and the ρ training sweep.

Stage	Device	Hours
Question writing	remote LLM	0.78
Figure captions	remote LLM	3.61
Multi-turn trajectories (train+eval)	remote LLM	6.49
Listwise LR	remote LLM	2.56
Prefix-freeze suffix-replan GAC	remote LLM	4.34
Earlier fixed-set GAC(not used in v5)	remote LLM	3.56
One full-stack BGE-M3 train	1 local GPU	0.09
Four v5 signal configs (sum)	local GPU	0.91
Six ρ trains (sum)	local GPU	2.63

F Additional Ablations and Diagnostics

F.1 Full transfer tables

Table 3 in the main text keeps one number per protocol. The complete metric sets are below.

Table 11: Retrieval on MMDocIR Eval (1,658 questions). Text retrievers use the same OCR pipeline as training (fair protocol). Jina-v4 encodes page images. Layout R@5 uses official parquet boxes.

Model	**Page R@1**	**Page R@5**	**Page R@10**	**nDCG@10**	**Layout R@5**
BGE-M3 (base)	0.418	0.705	0.812	0.585	**0.580**
Qwen3-Embedding-4B	0.448	0.734	0.822	0.609	0.432
Jina-embeddings-v4 (page image)	**0.574**	**0.848**	**0.907**	**0.729**	–
Hybrid (Qwen3 + Jina, RRF)	0.495	0.785	0.822	0.646	–
DocAgentic-R(ours)	0.420	0.710	0.812	0.589	0.571

Table 12: LongDocURL evidence-page retrieval (2,323 questions, 396 documents).

Configuration	**R@1**	**R@5**	**R@10**	**nDCG@10**	**MRR@10**
BGE-M3 (base)	0.477	0.762	0.847	0.585	0.599
LRonly	0.452	0.727	0.822	0.557	0.571
LR+GAC	0.491	0.759	0.847	0.588	0.606
LR+GAC+ECC	**0.494**	0.761	**0.851**	**0.592**	**0.609**
Full (+LOC)	0.492	**0.766**	0.851	0.592	0.608

The LongDocURL end-to-end table is Table 8 in the main text.

F.2 Page versus layout

On the fair protocol, the full model improves Page R@5 ($0.705 \rightarrow 0.710$) but has lower Layout R@5 than frozen BGE-M3 (0.571 vs 0.580). ColdStart examples emphasize long page text; BlueprintSynth LOClabels emphasize within-page blocks. This is a diagnostic, not a claim that LOCalready made localization uniformly better.

F.3 Evidence state

Matched $Q[\mathrm{SEP}]q_i$ remains the default.

Table 13: CMDR-Bench zero-shot page retrieval (800 questions, 255 documents).

Configuration	R@1	R@5	nDCG@5	R@10	nDCG@10
BGE-M3 (base)	0.161	0.440	0.309	0.555	0.346
LRonly	0.149	0.428	0.296	0.526	0.328
LR+GAC	0.171	**0.453**	0.320	**0.559**	0.354
LR+GAC+ECC	0.176	0.450	**0.320**	0.556	0.354
Full (+LOC)	**0.178**	0.446	0.319	0.559	**0.355**

Table 14: Turn-wise retrieval on BlueprintSynth-Eval v2 (2,500 questions, 7,686 turns). Not comparable to the 2,430-question static-Q main table.

Train	Eval	Chain R@20	Full@20	nDCG@10	MRR
BGE-M3	$Q[\mathrm{SEP}]q_i$	0.835	0.752	0.537	0.543
BGE-M3	$+ S_{i-1}$	0.782	0.694	0.484	0.495
Full	$Q[\mathrm{SEP}]q_i$	**0.848**	**0.768**	**0.555**	0.560
Full	$+ S_{i-1}$	0.786	0.697	0.495	0.508
Full (train $+ S$)	$Q[\mathrm{SEP}]q_i$	0.840	0.755	0.540	0.548
Full (train $+ S$)	$+ S_{i-1}$	0.835	0.747	0.551	**0.566**

F.4 Hard-negative inventory

F.5 Results by question type

Table 17 stratifies MMDocIR Page R@10 under the same fair OCR protocol as Table 11. Overall R@10 is unchanged at 0.812. The stratification does not support a type-specific headline.

F.6 Probes that do not change the operating point

A 200-question HotpotQA distractor probe (seed 20260728) answers 86% of questions from the full ten-paragraph pool, but only 25% of gold paragraphs change exact match when dropped, versus 0.8% of distractors. One frozen-agent $K{=}2$ cycle lowers Blueprint full-chain@20 from 0.674 to 0.660 and fills only 333 of 850 groups. Leave-one-out distillation of 4,861 training chains keeps 16.6% of gold units; 3,410 questions have $J{=}0$ on the full chain. An unseen subquery writer (Qwen3.8-27B, 200 questions) does not keep the chain gain: trajectory-union full-chain is 0.165 versus 0.140.

G Calibration Subset Protocol

Held-out tuning split. A 200-question calibration file (`calibration_ids.json`, seed 20260728, created 2026-08-11) holds 150 MMDocIR Eval questions stratified by domain and 50 LongDocURL questions stratified by question type. The file's usage rule is: checkpoint and hyperparameter selection only; never training; never a main-table subset. All numbered tables in this paper use the full evaluation sets.

What "calibration" does not mean. The same seed is the default RNG for training scripts. It does not denote a learned temperature, Platt scaling, or a per-modality score transform used in the main model. An experimental comparability head that added quantile-median biases and type embeddings *lowered* BlueprintSynth-Eval full-chain@20 from 0.674 to 0.659; a train/eval-matched $[MOD][GRAN]$ text prefix recovered 0.678. Neither extra module is used in the main tables (section 4.3). The retained comparability mechanism is the mixed-type InfoNCE denominator.

Table 15: Document-aware hard-negative taxonomy.

Negative type	**Construction**	**Forces the retriever to learn**
Same-doc adjacent page/chunk	same doc, page distance ≤ 2, off-chain	fine-grained discrimination
Cross-doc same-entity	same entity, different version/year	temporal/version awareness
Off-chain distractor	topically related, on no evidence chain	anti reward-hacking
Parsing-degraded pair (v1)	OCR text vs VLM caption of same page	parsing-noise awareness
CMDR-style hard negative	highly similar non-positive page	indirect-retrieval robustness

Table 16: Leave-one-negative-type ablation on BlueprintSynth-Eval static Q ($2{,}430$ questions).

Negatives	**Chain R@20**	**Full@20**	**nDCG@10**	**MRR**
Original (full v5)	0.833	0.674	0.681	0.800
All five types	0.835	0.676	0.687	0.808
Drop neighbor	0.836	0.677	0.683	0.801
Drop same-entity	0.837	0.680	0.685	0.807
Drop off-chain	0.839	0.683	0.683	0.802
Drop degrade pair	0.839	0.684	0.684	0.803
Drop CMDR-style	0.838	0.681	**0.697**	**0.821**

Table 17: MMDocIR Eval Page R@10 by question type, same OCR as training (fair protocol).

Question type	n	**BGE-M3**	**DocAgentic-R**	Δ
Figure	218	0.748	**0.766**	$+0.018$
Pure-text	193	0.834	**0.860**	$+0.026$
Table	173	**0.873**	0.867	-0.006
Chart	131	**0.885**	0.878	-0.007
meta-data	153	**0.680**	0.673	-0.007
text-only	383	**0.799**	0.796	-0.003
Generalized-text	62	**0.790**	0.726	-0.064
Overall	1658	0.812	0.812	0.000