MESA: Memory-Augmented Evidence Set Assembly for Setwise Reranking

Anonymous ACL submission

Abstract

Retrieval-augmented question answering depends on assembling a set of documents that *jointly* covers the information required to answer a query, yet existing rerankers score each candidate in isolation and leave this collective coverage to chance. We introduce **MESA** (Memory-Augmented Evidence Set Assembly), a setwise reranker that recasts top-k selection as a set-conditioned scoring problem in which each document is scored by its *marginal contribution* to the assembled evidence set. On an enterprise QA benchmark, MESA outperforms strong pointwise and cross-encoder rerankers on every ranking-quality metric, with the largest gains on human-curated queries that require multi-evidence aggregation. Controlled analysis supports the interpretation that learned cross-document interaction, rather than encoder capacity or post-hoc diversification, drives these gains, while a complementary root-cause study indicates that residual cross-style variability is best explained by training-distribution coverage. These results indicate that complementary evidence assembly is a learnable objective that pointwise rerankers lack any explicit mechanism to optimize.

1 Introduction

Retrieval-augmented generation (RAG) for question answering depends not on finding individually relevant documents but on assembling a set of documents that *collectively* contains the information needed to answer a query (Lewis et al., 2020; Cohen et al., 2025). When an answer requires evidence drawn from multiple distinct facets, k documents that each cover a different facet are far more useful to the downstream reader than k documents that repeat the same information, even when the latter score higher under any pointwise relevance criterion. The utility of a retrieved set is therefore properly measured by its *collective evidence coverage*: the degree to which the top-k documents jointly contain sufficient, non-redundant information to support the answer.

Existing rerankers lack a mechanism for marginal contribution. Bi-encoders (Reimers and Gurevych, 2019), cross-encoders (Nogueira and Cho, 2019; Nogueira et al., 2020), and large language model (LLM)-based listwise rankers (Sun et al., 2023) all assign a score to each candidate independently of the others and recover the top-k set as an artifact of sorted individual scores. No signal in this pipeline rewards a candidate for filling an evidence facet not yet covered by the assembled set, nor penalizes one for duplicating a facet the pool already contains. Post-hoc diversification heuristics such as Maximal Marginal Relevance (MMR) (Carbonell and Goldstein, 1998) attempt to repair this gap after scoring, but they apply redundancy suppression as a separate stage from relevance estimation rather than learning to optimize marginal evidence contribution end-to-end.

We introduce **MESA** (Memory-Augmented Evidence Set Assembly), a setwise reranker that recasts top-k selection as a set-conditioned scoring problem. Rather than scoring each candidate in isolation, MESA conditions each candidate's score on the full candidate pool, rewarding documents that fill uncovered evidence facets and penalizing those that duplicate what the pool already contains.

We report three findings. First, MESA improves ranking quality on WixQA, an enterprise RAG benchmark for answer-supporting evidence retrieval, with the largest gains on human-curated multi-facet queries where collective evidence coverage is the operative requirement. Second, controlled comparisons against same-encoder pointwise, post-hoc diversification, and generic setwise baselines narrow the source of the gain to learned cross-document interaction coupled with explicit coverage modeling, rather than to encoder capacity or post-hoc diversification. Third, zero-shot transfer to MuSiQue is limited, but MESA adapts

with as few as 99 target-domain queries, and the set-level gain persists under a 5× larger frozen backbone.

2 Related Work

Neural Reranking Without Cross-Document Reasoning. Cross-encoders (Nogueira and Cho, 2019; Nogueira et al., 2020) and late-interaction models (Khattab and Zaharia, 2020; Santhanam et al., 2022) substantially improve ranking quality by modeling rich query-document interactions, but they still score each candidate independently of what the rest of the retrieved set already contains. As a result, they can identify individually relevant documents without directly reasoning about whether the top-k set is collectively sufficient for answering the query.

Setwise Ranking Models. Setwise rerankers such as SetRank (Pang et al., 2020) and sequence-level rankers such as RankT5 (Zhuang et al., 2023) move beyond pointwise scoring by introducing cross-document context over the candidate pool, showing that ranking benefits from reasoning over a list or set rather than over isolated candidates. MESA is closest to this line of work, but differs in objective: it uses set-conditioned scoring not only to contextualize candidates, but also to make collective evidence coverage explicit through slot-based coverage tracking and a distinct novelty signal.

Diversification, Coverage, and Generalization. Another longstanding line of information retrieval (IR) research studies how ranked lists should cover multiple aspects of an information need rather than repeat the same evidence, from MMR (Carbonell and Goldstein, 1998) to intent-aware diversification methods such as IA-Select (Agrawal et al., 2009) and xQuAD (Santos et al., 2010). These approaches recognize the importance of coverage, but they typically operate as post-hoc selection rules on top of fixed relevance scores rather than as learned set-conditioned scorers. MESA extends this line by jointly supervising slot-based coverage and novelty alongside pointwise relevance, so the model learns to optimize complementary placement directly rather than repairing ranked lists after the fact.

3 MESA: Set-Conditioned Reranking

MESA scores each candidate by its marginal contribution to the full candidate pool, assembling top-k sets that jointly cover distinct evidence facets while suppressing redundant selections.

3.1 Problem Formulation

Pointwise rerankers score each candidate independently, leaving collective coverage of the assembled top-k set unmodeled. We instead formulate reranking as a set-conditioned scoring problem: given a query q, candidates $D=\{d_1, \ldots, d_n\}$ from a first-stage retriever, and ranks $\{r_i\}$, MESA learns

$$s_i = f(q,\, d_i,\, D_{\setminus i},\, r_i), \tag{1}$$

where $D_{\setminus i}$ denotes the remaining candidates, so each document's score reflects its *marginal contribution*—how much new, complementary information it adds beyond what the rest of the pool already covers.

Two complementary signals supervise this set-conditioned score: pointwise WixQA relevance labels define the ranking target for a listwise MLE loss ($\mathcal{L}_{\text{ListMLE}}$), while set-level binary cross-entropy supervision ($\mathcal{L}_{\text{SetBCE}}$) jointly penalizes false negatives and false positives over set membership in a single pass, distributing evidence coverage across the selected subset rather than concentrating scores on the single most relevant document. $\mathcal{L}_{\text{SetBCE}}$ serves as a proxy for collective coverage because WixQA's relevant-set labels are constructed so that each admitted document contributes information absent from the others.

3.2 Architecture: A Set-Conditioned Scoring Pipeline

Each stage of the pipeline addresses a distinct gap in pointwise scoring; Figure 1 illustrates the full design.

Input Representation. MESA takes a query q with BM25-returned candidates $D=\{d_1, \ldots, d_n\}$ and first-stage ranks $\{r_i\}_{i=1}^{n}$. Query and candidates are encoded by the same frozen MiniLM-L6-v2 (Reimers and Gurevych, 2019), yielding $\mathbf{e}_q \in \mathbb{R}^{384}$ and $\{\mathbf{e}_{d_i}\} \subset \mathbb{R}^{384}$; freezing stabilizes training on small datasets and permits offline caching of document embeddings. A query-aware projection lifts each candidate into the shared set-reasoning space,

$$\begin{aligned} \mathbf{h}_i^{(0)} &= W_{\text{doc}}\big[\mathbf{e}_{d_i};\, \mathbf{e}_q;\, \cos(\mathbf{e}_q, \mathbf{e}_{d_i})\big], \\ \mathbf{h}_q &= W_q\, \mathbf{e}_q, \end{aligned} \tag{2}$$

with $W_{\text{doc}} \in \mathbb{R}^{d\times 769}$, $W_q \in \mathbb{R}^{d\times 384}$, and d=256. This projection places every candidate in a shared

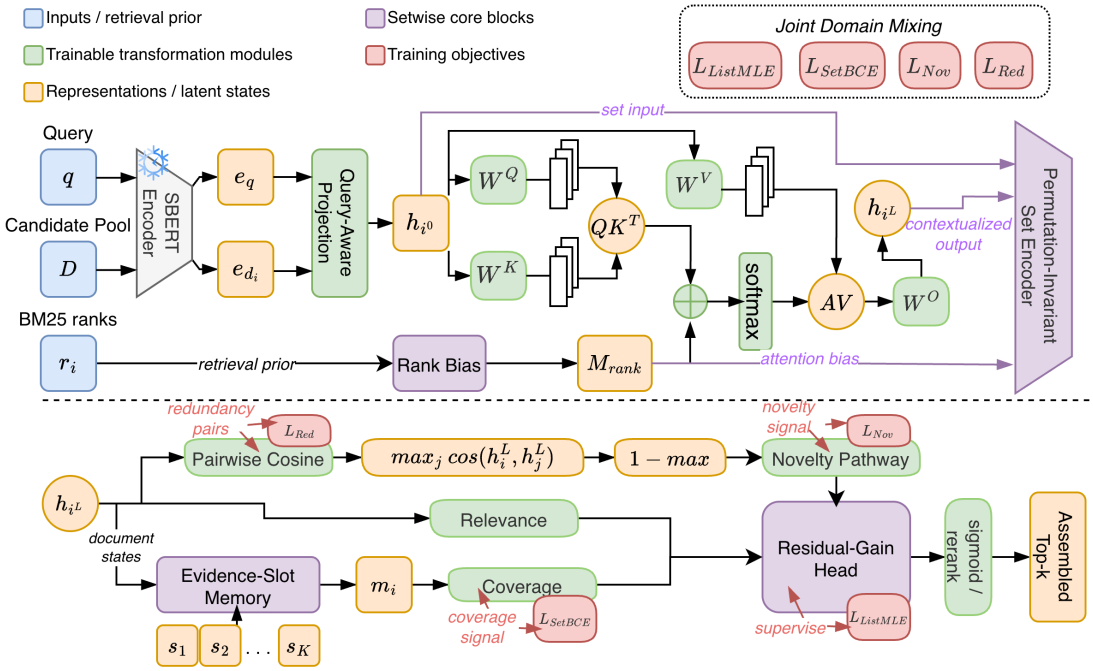

Figure 1: Overview of MESA. A frozen encoder produces query embedding $\mathbf{e}_q$ and document embeddings $\{\mathbf{e}_{d_i}\}$, which a query-aware projection maps to initial set states $\{\mathbf{h}_i^{(0)}\}$. BM25 ranks $\{r_i\}$ are preserved as a retrieval-side prior and injected into the set encoder as rank bias $\mathbf{M}_{\text{rank}}$. The permutation-invariant set encoder contextualizes each candidate against the full pool via rank-biased self-attention, yielding updated representations $\{\mathbf{h}_i^{(L)}\}$. These feed two parallel branches: the evidence-slot memory computes per-document slot-activation vectors $\mathbf{m}_i$, and the novelty pathway computes pairwise-cosine-based novelty scores n_i. The residual-gain head combines relevance, coverage, and novelty into the final score s_i, from which the assembled top-k evidence set is selected.

space that already encodes basic query relevance before cross-document reasoning begins. The appended cosine term exposes query-document compatibility directly, while the retrieval ranks are carried forward as a weak prior so that later stages can refine, rather than discard, first-stage evidence.

Permutation-Invariant Set Encoder. Cross-document interaction is necessary because a document's value depends not only on its relevance to the query but also on what the rest of the candidate pool already covers. The set encoder therefore processes $\{\mathbf{h}_i^{(0)}\}$ as an unordered set and turns each candidate into a pool-aware representation $\mathbf{h}_i^{(L)}$ whose score can reflect the surrounding evidence configuration rather than isolated relevance alone. To preserve first-stage retriever confidence without hard-coding the original ranking, the ranks $\{r_i\}$ are injected as a learnable pairwise attention bias before the softmax,

$$\mathbf{M}_{\text{rank}}[i,j] = \alpha \cdot \log\left(\frac{n+1-\min(r_i, r_j)}{n+1}\right), \tag{3}$$

where $\alpha \in \mathbb{R}$ is learned and the logarithmic form keeps the bias informative without allowing the initial retrieval order to dominate. The standard attention update then yields set-contextualized states $\mathbf{h}_i^{(L)} \in \mathbb{R}^d$.

Coverage and Novelty Branches. Coverage must be tracked separately from surface redundancy, because two documents can be topically similar yet contribute different evidence facets, or appear distinct while still filling the same underlying need. MESA therefore decomposes this question into two parallel signals computed from $\{\mathbf{h}_i^{(L)}\}$: an evidence-slot memory that exposes which latent evidence dimensions a document fills, and a novelty pathway that measures how distinctive the document remains relative to the pool. The *evidence-slot memory* maintains a bank of K=8 shared slot prototypes $\{\mathbf{u}_k\}_{k=1}^K \subset \mathbb{R}^d$ and produces per-document activations $m_{ik} = \sigma(\mathbf{h}_i^{(L)} \cdot \mathbf{u}_k) \in [0,1]$, collected in $\mathbf{m}_i \in [0,1]^K$. These activations make latent evidence coverage explicit to the scoring head: broad activation indicates that a document fills multiple uncovered dimen-

sions, whereas repeated activation on the same slots across candidates exposes redundancy that relevance alone cannot see. In parallel, the *novelty pathway* computes a differentiable MMR-style score $n_i = 1 - \max_{j \neq i} \cos\big(\mathbf{h}_i^{(L)}, \mathbf{h}_j^{(L)}\big)$. This term contributes a distinctiveness signal that is deliberately non-equivalent to slot coverage, allowing the scorer to distinguish between documents that cover new evidence and documents that are merely dissimilar on the surface.

Residual-Gain Head. The final score must combine the three ingredients that determine marginal contribution: relevance to the query, coverage of latent evidence needs, and distinctiveness from what the pool already contains. MESA realizes this objective with a residual-gain head that linearly combines these signals using learnable scalars w_{rel}, w_{slot}, $w_{\text{nov}} \in \mathbb{R}$,

$$s_i = w_{\text{rel}} \cdot \underbrace{\mathbf{h}_q \cdot \mathbf{h}_i^{(L)}}_{\text{relevance}} + w_{\text{slot}} \cdot \underbrace{\frac{1}{K} \sum_k m_{ik}}_{\text{coverage}} + w_{\text{nov}} \cdot \underbrace{n_i}_{\text{novelty}}, \tag{4}$$

where the query-document term contributes relevance, the mean slot activation contributes pool-aware coverage, and the novelty term suppresses redundant placements. Because $\mathbf{h}_i^{(L)}$ already encodes cross-document interaction, the resulting score measures each document's value relative to the assembled pool rather than in isolation.

3.3 Training: Joint Supervision for Ranking and Coverage

Jointly optimizing ranking quality and set-level coverage requires separate supervision signals for each: a listwise ranking loss rewards the correct ordering of relevant documents, while a set-level coverage loss penalizes assembled sets that miss relevant facets or include redundant ones. MESA combines these into a single training objective:

$$\begin{aligned} \mathcal{L} = {} & \lambda_{\text{rel}} \mathcal{L}_{\text{ListMLE}} + \lambda_{\text{set}} \mathcal{L}_{\text{SetBCE}} \\ & + \lambda_{\text{nov}} \mathcal{L}_{\text{Novelty}} + \lambda_{\text{red}} \mathcal{L}_{\text{Redundancy}}, \end{aligned} \tag{5}$$

with $(\lambda_{\text{rel}}, \lambda_{\text{set}}, \lambda_{\text{nov}}, \lambda_{\text{red}}) = (0.3, 0.3, 0.2, 0.2)$. $\mathcal{L}_{\text{ListMLE}}$ (Xia et al., 2008) optimizes permutation likelihood under pointwise relevance labels. $\mathcal{L}_{\text{SetBCE}}$ is binary cross-entropy over set-level relevance indicators, jointly penalizing false negatives and false positives across all candidates in a single pass. $\mathcal{L}_{\text{Novelty}}$ and $\mathcal{L}_{\text{Redundancy}}$ encourage high-scoring documents to be mutually dissimilar and penalize near-duplicate candidates that receive similar scores, respectively. The training mixture combines synthetic, expert-annotated, and simulated queries without domain-specific inference logic, exposing the coverage objective to the full range of query styles present at evaluation.

4 Experiments

4.1 Experimental Setup

Benchmarks. We evaluate on WixQA (Cohen et al., 2025), an enterprise RAG benchmark designed around evidence-assembly queries over a knowledge base of over 6,000 articles. Results are reported on two held-out test splits: *ExpertWritten*, comprising 200 human-curated queries that genuinely require multi-evidence aggregation, and *Simulated*, comprising 200 LLM-generated queries with lower lexical complexity. The two splits differ in query complexity and evidence-need depth, so performance across both tests whether coverage-aware reranking helps beyond the most demanding queries.

Baselines: Three Controls. Our comparison is scoped to the question of whether set-level inductive bias accounts for the observed gains, with each baseline isolating one alternative explanation rather than exhaustively covering the reranking landscape. BM25 serves as the first-stage sparse retriever. We then deploy three controls on top of it: (i) **Dense (MiniLM)**—same-encoder pointwise cosine, isolating MESA's set-level modules from the underlying encoder; (ii) **MMR-Dense** (λ=0.7, the best-performing value from a full λ sweep that confirms diversity forcing does not improve ranking quality on WixQA; details in Appendix F)—post-hoc diversity forcing on Dense scores, testing whether greedy diversification can account for MESA's gains; and (iii) **SetRank** (Pang et al., 2020)—learned permutation-invariant setwise attention with the same input representation and training mixture as MESA but replacing evidence-slot memory, novelty term, and rank bias with a simple linear head, isolating whether MESA's coverage decomposition contributes beyond generic setwise attention. **MonoT5-base** (Nogueira et al., 2020) is added as a strong cross-encoder reference at comparable model scale. LLM-based listwise rerankers (Sun et al., 2023) fall outside this scope since they reintroduce the model-scale confound that the Dense control is designed to rule out, and we leave their interaction with MESA's

Table 1: Main results on WixQA. **E**: ExpertWritten (200 queries); **S**: Simulated (200 queries). MESA: mean ± std over 3 seeds. Best per column in **bold**. Baseline scope and configurations are described in §4.

Method	Split	nDCG@5	nDCG@10	Recall@5	Recall@10	MRR@10
BM25	E	0.2977	0.3302	0.3983	0.4950	0.2916
	S	0.2385	0.2610	0.3475	0.4150	0.2176
Dense (MiniLM)	E	0.6357	0.6679	0.8017	0.8942	0.6067
	S	0.5665	0.6053	0.7575	0.8750	0.5247
MonoT5-base	E	0.6397	0.6700	0.8017	0.8892	0.6082
	S	0.6055	0.6333	0.7800	0.8650	0.5658
MMR-Dense (λ=0.7)	E	0.5777	0.6292	0.7017	0.8517	0.5754
	S	0.5083	0.5554	0.6650	0.8050	0.4847
SetRank	E	0.7205	0.7457	**0.8700**	**0.9450**	0.6908
	S	0.6210	0.6622	0.7750	**0.8975**	0.5946
MESA (joint)	E	**0.7408**±.012	**0.7646**±.014	0.8656±.012	0.9333±.019	**0.7241**±.012
	S	**0.6415**±.009	**0.6705**±.011	**0.7908**±.010	0.8783±.017	**0.6110**±.008

coverage objective to future work. BM25 retrieves 100 candidates for all rerankers, and MESA results throughout are reported as mean ± standard deviation across three random seeds.

4.2 Main Results

On WixQA, the benchmark that directly instantiates the evidence-assembly problem, MESA achieves the strongest ranking quality on both query styles, with its clearest advantage on the human-curated ExpertWritten split; Table 1 reports the full comparison. That split is the more diagnostic test: ExpertWritten queries require multi-evidence aggregation by construction, so collective evidence coverage—not pointwise relevance alone—is the decisive criterion.

The three controls show that this result cannot be reduced to any simpler explanation. Against Dense, the same-encoder control rules out the possibility that MESA merely benefits from stronger individual document representations. Against MMR-Dense, the post-hoc diversification control shows that naive novelty pressure is not enough and can actively misallocate the top ranks, so the gain cannot be explained as diversity alone. Against SetRank, the generic-setwise control shows that cross-document interaction by itself does not fully account for the result. The three controlled comparisons jointly narrow the source of MESA's advantage to set-level reasoning coupled with explicit coverage modeling; the same emphasis on complementary placement carries through to downstream answer quality, where MESA's coverage gain translates into more complete answers on multi-facet queries.

The comparison with SetRank also clarifies what the coverage-aware objective is optimizing. SetRank is strongest on Recall@10, but MESA leads on normalized discounted cumulative gain (nDCG) and mean reciprocal rank (MRR), especially near the top of the list where the reader actually consumes evidence. That trade-off is consistent with MESA's design: the goal is not to gather the largest possible number of relevant documents by rank 10, but to place complementary evidence earlier so that the top-ranked set is jointly more useful. This pattern supports a more specific claim than "setwise is better": coverage-aware reranking improves how evidence is arranged within the reader-visible ranks, even when another setwise model can recover slightly more relevant material at deeper cutoffs.

The remaining results delimit the conditions under which this main claim should be interpreted. On WixQA itself, a simple joint mixture of synthetic, expert, and simulated data yields the most stable improvement across both held-out splits, showing that the main result does not depend on domain tags or inference-time specialization. Beyond WixQA, the boundary conditions are organized by two axes: whether the evaluation criterion rewards complementary evidence, and whether the model has enough in-domain exposure to learn the relevant evidence patterns. MuSiQue is favorable on the first axis but not on the second in zero-shot transfer, so MESA does not beat Dense until balanced joint training supplies enough MuSiQue-specific supervision for the coverage objective to take effect. MESA does not fully transfer zero-shot to MuSiQue, but adapts with limited target-domain

Table 2: Component ablation (joint training, seed 42, 160/40 holdout). Δ relative to full model.

Variant	Exp	ΔE	Sim	ΔS
MESA (full)	.604	—	.538	—
Set-Level Module Removal				
w/o set encoder	.547	−.057	.528	−.010
w/o novelty	.547	−.057	.546	+.008
w/o slots	.595	−.009	.522	−.016
w/o rank bias	.559	−.045	.548	+.010
Reduced Variants (Toward Pure Pointwise)				
w/o query encoder	.593	−.012	.527	−.011
Rel-only†	.538	−.066	.520	−.018

† Set enc. + slots + novelty removed; pure pointwise.

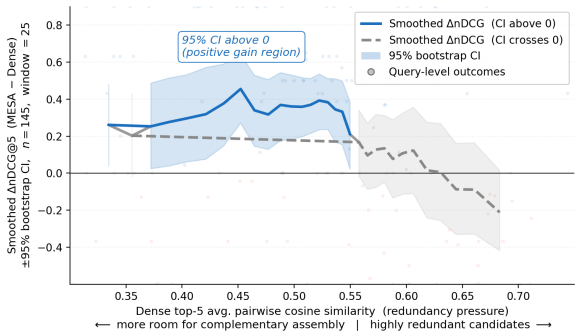

Figure 2: Quality gain of MESA over Dense (same-encoder control) across query-level redundancy pressure—measured by avg. pairwise cosine in Dense top-5—with smoothed per-query $\Delta\mathrm{nDCG@5}$ on the y-axis (95% bootstrap band; queries with `pool_n_rel` ≥ 1, n=145). MESA yields consistently positive gains in low-to-moderate redundancy regimes where the band remains above zero, and becomes uncertain only when Dense's top-5 is already highly redundant—consistent with MESA's complementary-evidence objective; the corresponding picture vs. MonoT5/SetRank is qualitatively similar but conflated with encoder and training-data confounds (§4.2).

supervision: with 25% of MuSiQue training data (99 queries), it matches or slightly surpasses the dense baseline (nDCG@5: 0.771 vs. 0.769), and improves further at full data; Appendix E provides the full sample-efficiency sweep underlying this interpretation. The same-encoder set-level gain also persists when the frozen backbone is replaced with BGE-base-en-v1.5 (109M), a 5× larger encoder, ruling out encoder capacity as the source of MESA's advantage; Appendix K reports this scale-up experiment. TREC DL is unfavorable on both axes: its precision-first criterion rewards early singleton relevance, and its supervision regime strongly favors models already aligned with passage-level relevance judgments. These checks therefore delimit the claim: MESA helps when reranking is genuinely an evidence-assembly problem, not whenever reranking is evaluated in the abstract.

5 Mechanism Analysis

Three questions organize this section: whether cross-document interaction drives the gain, whether the gain reflects evidence coverage rather than surface diversity, and what explains the residual gap across query styles.

5.1 Component Ablation

Cross-document interaction is the strongest observed contributor to MESA's gain on ExpertWritten. Removing the set encoder produces the clearest degradation on the human-curated split (−0.057 nDCG@5), establishing that MESA's main gain comes from evaluating documents in the context of one another rather than in isolation; the directional reading is robust across seeds, as the low cross-seed variance in Table 1 confirms, and Table 2 reports the full component breakdown. The asymmetry across query styles is equally informative: on Simulated, component contributions are smaller and less uniform, consistent with the weaker need for multi-facet evidence aggregation. The larger degradation of the Rel-only variant (−0.066) indicates that the slot and novelty branches contribute gain beyond cross-document interaction rather than restating it.

MESA's advantage concentrates where complementary placement is still possible. In low-to-moderate redundancy regimes—where Dense's top-5 still leaves room for additional evidence facets—MESA yields consistently positive per-query gains; uncertainty grows only when the pointwise ranking is already saturated with overlap, as Figure 2 visualizes across query-level redundancy pressure. This ties MESA's improvement to the regime its design targets: assembling complementary evidence sets rather than improving pointwise scores uniformly across all queries.

5.2 Coverage Analysis: Beyond Diversity

MESA's gains cannot be attributed to surface diversification; Table 3 rules out that account from three directions. MMR-Dense reduces intra-list similarity most aggressively yet underperforms Dense, showing that forced diversity can actively damage ranking quality. MonoT5 attains low intra-list simi-

Table 3: Ranking quality, redundancy, and slot coverage (ExpertWritten). Sim@5: avg. max-pairwise cosine in top-5 in MiniLM space (↓ = more diverse). Slots@1: avg. distinct slots above τ=0.3 at rank 1. SetRank Sim@5 and MESA nDCG@5 are 3-seed mean±std; remaining values are deterministic single-seed.

Method	**nDCG@5**↑	**Sim@5**↓	**Slots@1**↑
Dense (MiniLM)	0.636	0.723	1.22
MonoT5-base	0.640	0.508	—
SetRank	0.721	0.717 ±.004	—
MMR-D (λ=.5)	0.467	0.417	—
MMR-D (λ=.7)	0.578	0.616	—
MESA (joint)	**0.741** ±.012	0.673	**4.42**

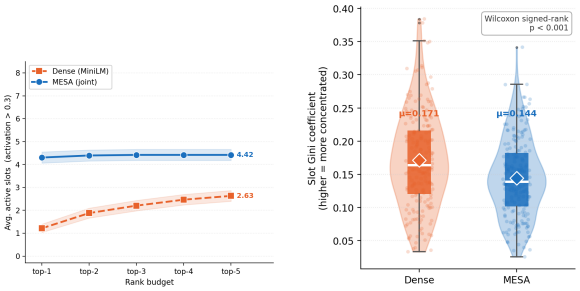

Figure 3: Aggregate slot coverage over 200 ExpertWritten queries. *Left:* Avg. distinct slots activated above τ=0.3 by rank budget; MESA activates 3.5× more slots at rank 1 (4.42 vs. 1.22) and maintains the gap across all budgets. *Right:* Per-query slot Gini coefficient; Dense (μ=0.171) concentrates activation on fewer slots than MESA (μ=0.144) in 78% of queries (Wilcoxon p<0.001), confirming that MESA distributes activation more evenly across latent evidence dimensions.

larity without materially surpassing Dense, indicating that surface similarity alone does not determine ranking quality. SetRank improves ranking quality without reducing surface redundancy, showing that learned cross-document interaction can help even without explicit diversification pressure. MESA differs from all three: it attains the strongest ranking quality while also exposing broader latent evidence coverage than Dense, not merely lower surface similarity. Active Slots@1 is reported only for the same-encoder pair because MonoT5 uses a different encoder and SetRank has no slot decomposition; under that controlled comparison, the combination of higher nDCG, intermediate redundancy, and broader slot activation points to purposeful evidence assembly rather than diversity alone.

MESA's advantage is accompanied by broader latent slot coverage across the ExpertWritten set. The gap opens at rank 1 and persists as the rank budget grows, and the slot distribution is more even

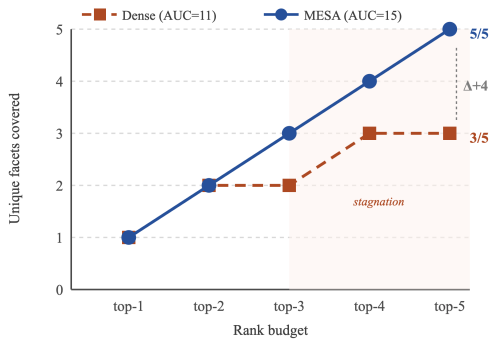

Figure 4: Cumulative human-audited facet coverage for a representative query whose answer requires five distinct evidence facets. MESA rises monotonically to full coverage by rank 5, whereas Dense plateaus at three of five facets after rank 2, illustrating how the aggregate slot-coverage gains in Figure 3 translate into broader coverage of manually labeled evidence dimensions.

for MESA than for Dense across most queries; Figure 3 plots both effects across all 200 ExpertWritten queries. The mechanism-level implication is direct: MESA improves ranking quality by covering more latent evidence dimensions per rank position, a property that surface redundancy metrics cannot observe.

The latent-slot advantage also maps onto human-interpretable evidence facets. On a representative five-facet query, MESA reaches full facet coverage within the top-5 whereas Dense stalls at three facets after rank 2; Figure 4 traces this cumulative coverage gap across rank positions. A manual audit by the authors across additional multi-facet WixQA queries confirms the same direction: across four audited queries, Dense covers 71% of annotated facets by rank 5 on average whereas MESA covers 94%, and on the broader annotation-derived multi-facet subset, SetCoverage@5 rises from 0.583 to 0.720, with the fraction of queries achieving full relevant-set coverage rising from 32.1% to 53.6%; Figure 4 traces the cumulative coverage gap for a representative five-facet query. This correspondence between latent slot activation and manually audited facet coverage supports the interpretation that broader slot coverage is not a latent-space artifact but reflects the answer-supporting evidence assembly the model is trained to optimize.

5.3 Root-Cause Analysis of the Expert–Simulated Gap

Among the factors we tested, training-distribution coverage best explains the Expert–Simulated gap. Neither Kullback–Leibler (KL) distillation from a

Table 4: Root-cause analysis for the Expert–Simulated gap (syn-only baseline). Δ relative to syn-only. Only in-domain data produces a meaningful gain.

Intervention	**ΔExp**	**ΔSim**	**Outcome**
KL distill. (λ=0.05)	$-.009$	$+.003$	× Rejected
KL distill. (λ=0.20)	$-.005$	$+.003$	× Rejected
Val on Sim split	$-.007$	$+.003$	× Rejected
Val on Exp+Sim	$\pm.000$	$\pm.000$	× Rejected
Train syn+sim†	$+.004$	$\mathbf{+.108}$	✓ **Confirmed**

† Upper bound: sim data in train *and* eval.

pointwise teacher nor changes to the early-stopping validation split materially improve Simulated performance; adding simulated queries to the training mixture yields the only substantial recovery on that split while leaving ExpertWritten largely intact, as Table 4 reports across all interventions. The implication is that the residual gap is better treated as a data-coverage problem than as evidence of a missing architectural mechanism: joint domain mixing is the most effective remedy among the interventions tested, and the gap narrows when the model sees the query style it will encounter at evaluation time.

The dominant downstream effect comes from the move to a stronger dense retrieval regime: once GPT-4o reads the top-k sets, most of the answer-quality jump is already explained by replacing sparse retrieval with dense retrieval. Within that stronger regime, the retrieval-side advantage created by evidence assembly carries through more weakly than it does in ranking metrics, but it is not washed out: cross-encoder reranking alone does not reproduce the same pattern, while MESA remains competitive with the strongest setwise result. Token-overlap metrics show only compressed downstream differences among strong rerankers; answer generation is a many-to-one mapping that partially smooths over near-equivalent evidence sets, so ranking improvements translate into smaller downstream margins than they do in nDCG or MRR. However, on multi-facet queries, where evidence assembly is the operative requirement, MESA yields the highest judged Comprehensiveness and Actionability among all evaluated methods, indicating that its coverage advantage translates into more complete answers precisely in the regime it is designed for; Appendix J reports the full judge evaluation including the prompt, scoring rubric, and per-method breakdown.

6 Conclusion and Discussion

Reranking for retrieval-augmented question answering is not always a problem of identifying the individually strongest passages. In settings where an answer depends on multiple complementary pieces of evidence, the operative objective is to assemble a top-ranked set whose members make distinct marginal contributions to answer support. Three controlled comparisons—against same-encoder Dense, MMR-Dense, and SetRank—support this interpretation for MESA: each rules out one alternative explanation, and their joint evidence narrows the source of the gain to set-level reasoning with explicit coverage modeling, rather than to a stronger encoder or post-hoc diversification.

The boundary conditions on this claim are equally clear. MESA's advantage concentrates in low-to-moderate redundancy regimes where the pointwise top-5 still leaves room for complementary placement; mechanism-level analyses link the gain to broader latent and manually audited evidence coverage, not to surface diversity alone. The Expert–Simulated gap is better treated as a data-coverage problem than as evidence of a missing architectural mechanism: joint domain mixing recovers most of the gap when the model encounters the query style present at evaluation. In downstream generation, token-level metrics compress differences among strong rerankers, consistent with the many-to-one mapping from evidence sets to generated answers; on multi-facet queries, where evidence assembly is the operative requirement, MESA achieves higher judged comprehensiveness and actionability than pointwise and cross-encoder baselines.

A conditional design implication follows: when answer quality depends on what the top-ranked passages *collectively* cover rather than on how each scores individually, a relevance-only objective leaves a systematic gap that set-level reasoning can close. This condition is operative when queries require multiple complementary pieces of evidence, when early rank positions determine what the reader consumes, and when pointwise rankings leave room for non-redundant coverage gains. The benefit is not universal: precision-first benchmarks and zero-shot cross-domain transfer both delineate the boundary beyond which set-level objectives do not improve over relevance-only scoring.

Limitations

MESA's boundaries are organized by three dimensions: the scope of the coverage objective, the strength of the mechanistic evidence, and the data and encoder scales at which the claims have been verified.

Scope of the Coverage Objective. MESA's coverage objective aligns naturally with multi-evidence QA but creates a structural tension with precision-first benchmarks such as TREC DL, where graded relevance rewards early precision over evidence coverage and domain-specific pre-training provides a strong prior; the cross-domain results in Section 4.2 show that small-scale external benchmarks additionally introduce validation-protocol sensitivity. Slot-based coverage alone does not capture multi-step reasoning chains without task-specific signal: the evidence-slot memory tracks which latent dimensions a document fills but does not model the logical dependencies between evidence pieces that multi-hop queries require. Evidence slots are also re-initialized per query at inference time, leaving the persistent slot state envisioned for update-aware retrieval settings as future work.

Strength of Mechanistic Evidence. Our mechanistic analyses are narrower in scope than the main benchmarking. The component ablation is reported for a single representative seed, and the root-cause study examines a limited set of interventions selected to discriminate among the candidate explanations we considered most plausible; Table 2 and Table 4 report the full breakdowns. These results therefore support the proposed interpretation that cross-document interaction drives the ExpertWritten gain and that training-distribution coverage best accounts for the Expert–Simulated gap, but they do not fully establish a unique causal account: alternative architectural choices we did not test or training-distribution interventions outside our intervention set could in principle reproduce parts of the observed pattern. Multi-seed ablations and a broader intervention space, including counterfactual data ablations and additional teacher-signal designs, would strengthen the causal claims.

Scale. Joint training uses 6,541 total queries. Whether MESA's set-level inductive bias yields similar advantages at larger data scales—where stronger pointwise baselines may narrow the gap through sheer coverage—remains unverified. The set-level advantage is not diminished when the frozen backbone is replaced with a 5× larger encoder (BGE-base-en-v1.5, 109M), as the encoder scale-up experiment reported in Appendix K confirms, but data-scale robustness remains an open question. On domain coverage, MESA does not establish strong zero-shot generalization: zero-shot transfer to MuSiQue trails Dense, and the adaptation signal at 25% target-domain data is small in absolute terms (+0.002 nDCG@5 over Dense); Appendix E provides the full sample-efficiency sweep underlying this boundary.

Potential Risks. Coverage-oriented reranking can increase the perceived completeness of assembled evidence sets, which introduces operational risks that depend on corpus quality and deployment context. When the underlying corpus is incomplete or outdated, a more coherent top-k evidence set may give users a stronger impression of answer completeness than the corpus actually supports, increasing the risk of over-trust in system outputs. Similarly, corpus-level biases or systematic facet omissions are not corrected by the coverage objective; rather, they may be assembled more persuasively than they would be under pointwise ranking. In enterprise deployments with weak access controls, stronger evidence assembly may also make it easier to surface sensitive or access-restricted content than a relevance-only reranker. These risks are best mitigated through permission-aware retrieval, corpus governance, provenance-aware interfaces that expose the sources underlying each assembled evidence set, and human oversight in high-stakes decision settings.

References

Rakesh Agrawal, Sreenivas Gollapudi, Alan Halverson, and Samuel Leong. 2009. Diversifying search results. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pages 5–14.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, and 1 others. 2016. MS MARCO: A human generated machine reading comprehension dataset. In *Proceedings of the Workshop on Cognitive Computation: Integrating Neural and Symbolic Approaches*.

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, pages 335–336. ACM.

Dvir Cohen, Lin Burg, Sviatoslav Pykhnivskyi, Hagit Gur, Stanislav Kovynov, Olga Atzmon, and Gilad Barkan. 2025. Wixqa: A multi-dataset benchmark for enterprise retrieval-augmented generation. *arXiv preprint arXiv:2505.08643*.

Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen M. Voorhees. 2020. Overview of the TREC 2019 deep learning track. In *Proceedings of the 28th Text REtrieval Conference*.

Jiajie Jin, Yutao Zhu, Zhicheng Dou, Guanting Dong, Xinyu Yang, Chenghao Zhang, Tong Zhao, Zhao Yang, and Ji-Rong Wen. 2025. Flashrag: A modular toolkit for efficient retrieval-augmented generation research. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 737–740.

Omar Khattab and Matei Zaharia. 2020. ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 39–48.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with BERT. *arXiv preprint arXiv:1901.04085*.

Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. 2020. Document ranking with a pre-trained sequence-to-sequence model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 708–718.

OpenAI. 2024. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Liang Pang, Jun Xu, Qingyao Ai, Yanyan Lan, Xueqi Cheng, and Jirong Wen. 2020. SetRank: Learning a permutation-invariant ranking model from information retrieval. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 499–508.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2022. ColBERTv2: Effective and efficient retrieval via lightweight late interaction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3715–3734.

Rodrygo L. T. Santos, Craig Macdonald, and Iadh Ounis. 2010. Exploiting query reformulations for web search result diversification. In *Proceedings of the 19th International Conference on World Wide Web*, pages 881–890.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023. Is ChatGPT good at search? investigating large language models as re-ranking agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14918–14937.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. MuSiQue: Multihop questions via single-hop question composition. In *Transactions of the Association for Computational Linguistics*, volume 10, pages 539–554.

Fen Xia, Tie-Yan Liu, Jue Wang, Wensheng Zhang, and Hang Li. 2008. Listwise approach to learning to rank: Theory and algorithm. In *Proceedings of the 25th International Conference on Machine Learning*, pages 1192–1199.

Honglei Zhuang, Zhen Qin, Rolf Jagerman, Kai Hui, Ji Ma, Jing Lu, Jianmo Ni, Xuanhui Wang, and Michael Bendersky. 2023. RankT5: Fine-tuning T5 for text ranking with ranking losses. *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2308–2313.

A Case Study: Evidence Facet Coverage

Manual facet analysis by the authors on four representative multi-evidence WixQA queries grounds the latent-slot interpretation from the main paper in human-interpretable evidence. This analysis is a manual inspection conducted by the authors over existing public benchmark examples; we did not recruit external annotators or run a human-subject study. Table A1 summarizes per-query facet coverage at rank budget 5; Appendix B provides the per-document visualization for Case D, and Figure 4 in the main paper traces the cumulative coverage curve for the same query.

MESA achieves full facet coverage in the majority of audited queries while Dense consistently plateaus before exhausting the available facets, a pattern directionally consistent with the latent-slot analysis in Figure 3.

Table A1: Manually audited facet coverage at rank budget 5 across four multi-facet WixQA queries (author inspection over existing benchmark examples). J: number of annotated facets; entries show unique facets covered out of J. MESA covers 94% of annotated facets on average vs. 71% for Dense (+23%pp). †Equal final count, but MESA covers the query-critical verification facet (F1) while Dense covers a peripheral SEO facet (F4).

Case	Query	J	Dense	MESA
A (q14)	Guest checkout	4	3/4 (75%)	4/4 (100%)
B (q171)	GSC verify†	4	3/4 (75%)	3/4 (75%)
C (q185)	Video autoplay	4	3/4 (75%)	4/4 (100%)
D (q177)	Mobile overlap	5	3/5 (60%)	5/5 (100%)

We extend this manual-analysis picture to the full annotation-derived relevant set with the SetCoverage@5 measure, defined as $|\text{top-5} \cap \text{Rel}(q)|/|\text{Rel}(q)|$. SetCoverage@5 is identical to Recall@5 when computed over the annotated relevant set, so the all-query row in Table A2 is reported for completeness. The more diagnostic test is the multi-facet subset ($|\text{Rel}| \geq 2$), which isolates queries where collective evidence assembly can produce a meaningful difference in coverage.

Table A2: Annotation-derived SetCoverage@5 (= Recall@5 on the annotated relevant set) on WixQA ExpertWritten. *FullCov*: fraction of queries where all relevant documents are covered by the top-5. *Win/Tie/Loss*: per-query comparison (MESA vs. Dense). The multi-facet subset ($|\text{Rel}| \geq 2$, N=28) is the more diagnostic test; all-query figures are reported for completeness.

Subset	N	SetCov@5		FullCov	
		Dense	MESA	Dense	MESA
All	200	0.802	0.816	76.5%	79.0%
$\|\text{Rel}\|\geq 2$†	28	0.583	0.720	32.1%	53.6%

†Multi-facet subset; Win/Tie/Loss: 18/8/2 (MESA).

B Aggregate Slot Analysis

Two complementary views extend the aggregate slot-coverage analysis in Figure 3: an auxiliary diagnostic of mean slot activation magnitude per rank position (Figure A1), and the per-document case-study visualization for Case D (Figure A2). The principal coverage evidence remains the *number of distinct slots activated* in Figure 3; activation magnitude is reported here only as a calibration check on the underlying magnitudes, not as a quality signal in its own right.

The activation-magnitude gap is largest at rank 1 and decays monotonically, confirming that the per-rank distinct-slot counts in Figure 3 are not driven by Dense activations sitting just below the τ=0.3 threshold.

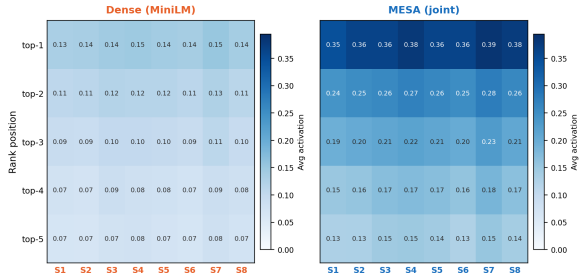

Figure A1: *Auxiliary diagnostic.* Mean slot activation magnitude per rank position averaged over 200 ExpertWritten queries. MESA's top-1 documents activate slots at ≈0.35–0.39, roughly 3× higher than Dense (≈0.13–0.15); both methods decay consistently with rank, and the uniform pattern across slots is consistent with a shared evidence-need decomposition learned across the training distribution. Higher activation magnitude is not itself a quality criterion: a single document could in principle activate one slot strongly, while MESA's coverage advantage in Figure 3 comes from activating *more distinct* slots above τ=0.3. This figure is therefore intended as a calibration check on the magnitudes underlying the slot-count metric, not as direct evidence of ranking quality.

Tracing Case D through MESA's four pipeline stages makes the data flow concrete and connects the architecture to the coverage outcome. *Stage 1 (Input Representation):* BM25 returns 100 candidates with ranks $\{r_i\}$; the projection of Eq. 2 lifts each candidate into the shared set-reasoning space $\mathbf{h}_i^{(0)}$. *Stage 2 (Set Encoder):* rank-biased self-attention (Eq. 3) updates each $\mathbf{h}_i^{(0)}$ into a set-contextualized $\mathbf{h}_i^{(L)}$, so that a candidate covering a facet already addressed by a high-confidence neighbor is down-weighted before scoring. *Stage 3 (Coverage and Novelty Branches):* the slot-activation vectors $\mathbf{m}_i$ for the two ground-truth relevant documents fire on complementary slots, while the novelty score n_i remains low for near-duplicate candidates further down the pool. *Stage 4 (Residual-Gain Head):* Eq. 4 combines the three signals so that the two relevant documents occupy ranks 1–2, and each successive rank introduces a new evidence facet (Figure A2), achieving full coverage (5/5) by rank 5. The same query under a pointwise Dense baseline retrieves zero relevant documents in the top-5 and covers only 3/5 facets, stagnating after rank 2.

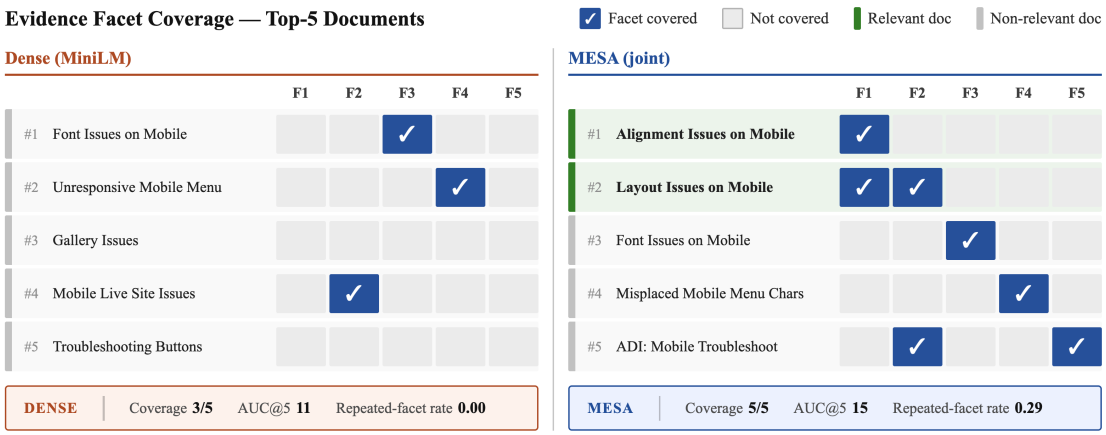

Figure A2: Top-5 document sets selected by Dense (left) and MESA (right) on the representative WixQA query Case D (*"Elements are misplaced and overlapping on mobile site view"*). Each row is one document; columns are five manually labeled evidence facets. Dense retrieves zero relevant documents and covers only 3 of 5 facets, stagnating after rank 2. MESA places both relevant documents in the top-2 and introduces a new facet at each successive rank, achieving full coverage (5/5) by rank 5.

C Domain Training Progression

Joint training without domain tags yields the most stable simultaneous improvement across both WixQA splits; Table A3 reports the three-seed recipe comparison and Table A4 shows the underlying single-seed mix-ratio progression. Mix ρ denotes the fraction of simulated queries added to synthetic training, and Tag-$*$ variants prepend a domain tag to each input.

Table A3: Training recipe comparison on WixQA (160/40 holdout; 3-seed mean ± std). Joint training without domain tags improves both splits simultaneously and yields the lowest Simulated variance.

Recipe	Tag	Exp nDCG@5	Sim nDCG@5
Syn only (base)	×	.541±.015	.496±.020
Joint (all)	×	**.563**±.025	**.525**±.014
Tag-train	✓	.547±.022	.519±.031
Tag-all	✓	.570±.025	.521±.006

Table A3 shows that the joint recipe is the only configuration that improves both splits simultaneously while maintaining low variance; tag-based variants either widen the Expert–Simulated gap or introduce higher seed-to-seed instability.

Table A4 shows that simple ratio mixing improves Simulated monotonically but leaves ExpertWritten flat, indicating that data quantity alone cannot overcome distributional mismatch; the joint recipe resolves this tension and is therefore selected as the main training configuration.

Table A4: Domain training progression across mix ratios (seed 42, 160/40 holdout, nDCG@5). Mix: fraction of simulated queries added to synthetic training. Simple ratio mixing improves Simulated but not ExpertWritten; the joint recipe achieves the best ExpertWritten while maintaining Simulated gains.

Variant	Mix	Expert	Sim	Gap
Baseline (syn)	0.00	.557	.490	−.067
Mix 0.10	0.10	.568	.499	−.069
Mix 0.25	0.25	.567	.517	−.051
Mix 0.50	0.50	.558	.504	−.054
Mix 1.00	1.00	.554	.524	−.030
Joint (all)	all	**.592**	**.543**	−.049
Tag-train (syn+sim)	syn+sim	.564	.539	−.025
Tag-all (all)	all	.603	.510	−.093

D External Benchmark Evaluation

MuSiQue and TREC DL reveal complementary boundary conditions for MESA's set-level inductive bias: the former tests structural alignment without in-domain exposure, the latter tests precision-first criteria misaligned with coverage objectives. Table A5 summarizes the headline comparison; the sections below break out zero-shot transfer numbers and the validation–test mismatch that produced an earlier biased TREC DL estimate. The sample-efficiency sweep underlying the MuSiQue adaptation result is in Appendix E.

The two settings reveal complementary boundary conditions for MESA's set-level inductive bias. On MuSiQue—structurally aligned with multi-facet coverage—zero-shot MESA trails Dense, confirming that the inductive bias does not transfer

Table A5: Cross-domain nDCG@5 summary. MuSiQue (Trivedi et al., 2022): balanced joint (WixQA+2×MuSiQue) from WixQA-pretrained init. TREC DL (Craswell et al., 2020): year-based protocol (train DL'20, evaluate DL'19, $N = 43$).

Method	MuSiQue	TREC DL
Dense (MiniLM)	.769	.867
MonoT5-base	.859	**.898**
MESA (zero-shot)	.726	.831
MESA (joint)	**.796**±.010	.835±.015

without in-domain exposure; balanced joint training with MuSiQue oversampled at roughly 2.5× the WixQA query count lifts MESA above Dense, while a less aggressive ratio yields negligible gain, indicating that MuSiQue evidence patterns must dominate the learning signal for the coverage objective to take effect. On TREC DL—a precision-first criterion misaligned with set-level coverage—MESA (joint) remains competitive with Dense but trails MonoT5-base; the gap reflects MonoT5's direct MS MARCO (Bajaj et al., 2016) supervision aligning tightly with TREC's passage-level relevance judgments, compounded by an evaluation criterion that does not reward complementary evidence assembly. MuSiQue and TREC DL jointly delineate the conditions under which set-level optimization confers an advantage, rather than supporting a general transferability claim.

Table A6 and Table A7 report the full zero-shot and joint-training comparisons across both benchmarks; balanced joint training recovers TREC DL 2019 performance under a corrected year-based protocol, compared to the biased estimate analyzed below.

Table A6: Zero-shot transfer to TREC DL 2019 (Craswell et al., 2020) (MS MARCO passage reranking, 43 queries, WixQA synthetic-only training). MESA trails MonoT5-base by 6.7pp nDCG@5, reflecting TREC's precision-first criterion and MonoT5's direct MS MARCO supervision.

Method	nDCG@5	nDCG@10	MRR@10
Dense (MiniLM)	0.867	0.847	0.936
MonoT5-base	**0.898**	**0.886**	**0.968**
MESA	0.831	0.812	0.895

Table A8 tracks the val–test nDCG@5 gap across training epochs for the original joint-training protocol (representative seed). The gap grows monotonically, reaching +0.30 by epoch 15, confirming that the 10-query validation slice overstates test generalization throughout training. This is a *selection artifact*: multi-domain early stopping preferentially selects checkpoints that overfit the easy TREC validation signal. An earlier joint-training estimate of 0.576 on TREC DL was a direct artifact of this design. A corrected year-based protocol (train on DL'20 with a 40-query WixQA held-out validation set, evaluate on full DL'19) eliminates the artifact and recovers 0.835±0.015 (three seeds), the value reported in Table A5.

Table A7: Zero-shot transfer to multi-hop benchmarks (WixQA synthetic-only training). MESA trails Dense on both corpora zero-shot, confirming that the coverage inductive bias requires in-domain exposure. After balanced joint training (WixQA+2×MuSiQue), MESA reaches MuSiQue nDCG@5 0.796±0.010, exceeding Dense by +2.7pp; joint results are in Table A5.

Corpus	Method	nDCG@5	R@5	MRR@10
2WikiMHQA	Dense	.840	.915	.858
	MonoT5	**.917**	**.968**	**.920**
	MESA	.784	.883	.799
MuSiQue	Dense	.769	.816	.883
	MonoT5	**.859**	**.891**	**.939**
	MESA	.726	.785	.844

Table A8: TREC DL validation–test nDCG@5 gap under joint training (seed 42). The monotonically growing gap identifies a selection artifact: the 10-query validation slice preferentially selects checkpoints that overfit TREC, overstating test generalization throughout training.

Epoch	Val nDCG@5	Test nDCG@5	Gap
1	0.658	0.562	+0.096
3	0.843	0.648	+0.195
5	0.865	0.622	+0.243
8	0.896	0.643	+0.253
10	0.902	0.634	+0.268
15	0.926	0.624	+0.303

E Sample Efficiency on MuSiQue

This appendix reports a sample-efficiency sweep that characterizes how quickly MESA's set-level mechanism adapts to an unseen multi-hop QA domain when initialized from WixQA-pretrained weights. Starting from a WixQA-pretrained checkpoint, we jointly train MESA with varying fractions (5%–100%) of MuSiQue training data mixed with the fixed WixQA anchor (320 expert+simulated queries) and evaluate on a held-out MuSiQue test set (100 queries). The WixQA anchor is never re-

placed; MuSiQue queries are oversampled so that at 100% they constitute roughly 2.5× the WixQA query count, consistent with the balanced joint recipe in Appendix D. All 15 runs (5 fractions × 3 seeds) share the same pretrained initialization (WixQA seed 42, epoch 8) to isolate the effect of target-domain data quantity.

Table A9: Sample-efficiency sweep on MuSiQue (multi-hop QA). MESA starts from a WixQA-pretrained checkpoint and is jointly fine-tuned with increasing fractions of MuSiQue data alongside a fixed WixQA anchor (320 queries); nDCG@5 averaged over 3 seeds (±std). Zero-shot MESA trails Dense by 4.3pp; the crossover occurs at 25% (99 queries), and performance improves monotonically through full data.

Fraction	#Queries	nDCG@5	Δ vs Dense
Dense (MiniLM)	—	0.769	—
0% (zero-shot)	0	0.726	−0.043
5%	19	0.767±.007	−0.002
10%	39	0.754±.005	−0.015
25%	**99**	**0.771**±.004	**+0.002**
50%	198	0.769±.003	+0.000
100%	396	0.785±.005	+0.016
MonoT5-base	—	0.859	+0.090

The crossover occurs at **25%** (99 queries), where MESA first surpasses the dense retrieval baseline; performance then improves monotonically through full data, consistent with the joint-training result in Table A5. The non-monotonicity at 10% lies within the noise range given the reported standard deviations and likely reflects the high variance characteristic of small training fractions. WixQA test performance is maintained across all fractions (nDCG@5 ≈ 0.639–0.656), confirming that joint training does not degrade source-domain performance.

These results do not establish strong zero-shot generalization; rather, they show that MESA's set-level mechanism can be adapted to a structurally different multi-hop QA domain with limited target-domain supervision, suggesting partial transferability of the pretrained coverage inductive bias.

F MMR-Dense λ Sweep

All λ values in the MMR-Dense sweep fall below Dense on nDCG@5, confirming that post-hoc diversity forcing does not improve ranking quality on WixQA; the best-performing value (λ=0.7) is used as the MMR-Dense baseline in Table 1. Smaller λ places more weight on the diversity penalty, while larger λ places more weight on relevance.

Across all λ values, MMR-Dense remains below Dense on nDCG@5. Performance improves monotonically with larger λ, indicating that the diversity penalty is uniformly harmful for ranking quality on WixQA and that the test queries benefit from high-relevance evidence concentration rather than forced lexical diversity.

G Evaluation Metrics

We report five complementary IR metrics across all experiments. Normalized Discounted Cumulative Gain (nDCG@k) measures ranking quality by rewarding relevant documents placed higher in the list, weighted by position and normalized against the ideal ranking; we report it at k=5 and k=10. Recall@k measures the proportion of all relevant documents in the corpus that appear in the top-k results, capturing coverage rather than ordering. Mean Reciprocal Rank (MRR@10) is the average of the reciprocal rank of the first relevant document across queries, reflecting how quickly each method surfaces at least one useful result. The three metrics are complementary: nDCG captures graded ranking quality, Recall captures set-level coverage, and MRR captures early precision, providing a multi-faceted view of retrieval utility for downstream RAG use.

H Training Details

Optimization and Architecture. MESA totals approximately 2.5M trainable parameters and is trained for 15 epochs with early stopping (patience 4) using AdamW (weight decay 0.01) with a one-cycle learning rate schedule (peak LR 8×10^{-5}, warmup 10%), batch size 16, on an A10G GPU. The model uses a hidden dimension of d=256, a two-layer set encoder with four attention heads, K=8 evidence slots, and a maximum of 100 can-

Table A10: MMR-Dense λ sweep on WixQA test splits. Every tested λ value underperforms Dense on nDCG@5, and the diversity penalty is monotonically harmful: ranking quality improves as diversity pressure decreases. Best-performing λ=0.7 is used as the MMR-Dense baseline.

	Expert			Sim		
λ	nDCG@5	R@5	MRR	nDCG@5	R@5	MRR
0.3	.427	.425	.451	.349	.355	.358
0.5	.467	.507	.487	.388	.440	.399
0.7	.578	.702	.575	.508	.665	.485
Dense	.636	.802	.607	.567	.758	.525

didates per query. The joint training mixture comprises 6,221 synthetic queries, 160 expert-annotated queries, and 160 simulated queries from WixQA (Cohen et al., 2025). Loss weights are fixed at $(\lambda_{\text{rel}}, \lambda_{\text{set}}, \lambda_{\text{nov}}, \lambda_{\text{red}}) = (0.3, 0.3, 0.2, 0.2)$ throughout all experiments.

Learned Scalar Weights. The rank bias scalar α (Eq. 3) is initialized to 1.0 and learned without constraints. Its converged value is $\alpha \approx 1.25$ (mean across three seeds, std < 0.01), indicating that the model actively leverages retrieval-rank confidence rather than discounting it. This is consistent with the observation that BM25 rank is a reliable prior for human-curated queries (§5.1). The residual-gain head scalars $w_{\text{rel}}, w_{\text{slot}}, w_{\text{nov}}$ converge to $\approx$0.48, $\approx$0.33, $\approx$0.16 respectively (std < 0.01 across seeds), and all three are strictly positive, confirming that coverage and novelty contribute meaningfully alongside relevance. The relative proportions ($\approx$3:2:1) reflect the intended design: relevance is the primary signal, with coverage and novelty providing structured secondary corrections toward complementary evidence assembly.

Coverage-Analysis Threshold. The slot activation threshold τ=0.3 used in coverage analysis (§5.2) was set as a conservative midpoint of the sigmoid output range, and results are stable across $\tau \in [0.2, 0.4]$.

I Architecture Sweep: E1–E4

This appendix complements the ablation in §5.1 with an architectural sweep over four candidate extensions to MESA: bi-state encoding (E1), an explicit marginal-gain head (E2), field-aware projections (E3), and replay augmentation (E4). Each row enables a different combination of these extensions on top of the base MESA architecture.

Table A11: Architecture sweep (3-seed mean±std, 200-query eval). E1: Bi-state, E2: Marginal-gain, E3: Field-aware, E4: Replay aug. ⋆: proposed MESA.

Variant	E1	E2	E3	E4	Exp	Sim	Param
MESA⋆	×	×	×	×	**.741**±.012	.642±.009	2.5M
+E1	✓	×	×	×	.728±.009	**.643**±.010	3.0M
+E2	×	✓	×	×	.729±.014	.625±.012	3.0M
+E1+E2	✓	✓	×	×	.738±.006	.621±.003	3.6M
+E3	×	×	✓	×	.729±.007	.642±.005	2.8M
+E4	×	×	×	✓	.742±.013	.635±.011	2.5M
+E3+E4	×	×	✓	✓	.728±.014	.642±.026	2.8M
+All (E1–E4)	✓	✓	✓	✓	.737±.019	.639±.015	4.5M
+3-Layer	✓	✓	×	×	.698±.008	.592±.009	3.8M

All E1–E4 extensions stay within ±0.02 of MESA on ExpertWritten, and none consistently exceed MESA on both splits simultaneously. The 3-layer variant shows the clearest overfitting signal. These results corroborate the main ablation (§5.1): MESA's compact 2-layer design with K=8 slots is matched to the current data scale, and additional capacity tends to overfit rather than generalize.

J Downstream RAG Evaluation

The dominant downstream effect is the shift from sparse to dense retrieval; setwise rerankers add a smaller but non-negligible increment, and MESA's clearest advantage appears on multi-facet queries under judge-based evaluation. We evaluate end-task generation by feeding each method's top-5 documents to GPT-4o (OpenAI, 2024) under a shared customer-support prompt template, identical decoding settings (temperature 0, max 1024 tokens), and the same 200-query splits used throughout the paper. Generated answers are scored against gold WixQA answers using token-level F1 and Recall-Oriented Understudy for Gisting Evaluation (ROUGE-1) (Jin et al., 2025) after standard normalization (lowercasing, article removal, punctuation stripping). All methods share the BM25 candidate pool, so any difference in output quality is attributable to how the top-5 evidence set is assembled rather than to retrieval coverage or generation capacity.

Table A12: Downstream generation quality (GPT-4o reader, top-5 evidence per method, token-level F1 and ROUGE-1, 200 queries per split). The largest quality jump corresponds to the shift from sparse to dense retrieval; setwise methods add a smaller increment on top.

	Expert		Sim	
Method	**F1**	**R-1**	**F1**	**R-1**
BM25	.519	.400	.421	.369
Dense (MiniLM)	.619	.491	.485	.398
MonoT5-base	.520	.401	.421	.367
SetRank	**.622**	**.494**	**.506**	**.415**
MESA (joint)	.614	.487	**.506**	.413

A clear pattern emerges across the pointwise-to-setwise progression. BM25 and MonoT5-base form a tight cluster (F1 $\approx$ .520/.421), suggesting that cross-encoder reranking alone does not reliably improve the answer-supportive quality of the assembled top-5. Dense retrieval (MiniLM cosine) breaks this plateau substantially (Expert F1 +0.100, Sim F1 +0.064), consistent with its higher nDCG@5 in Table 1 and indicating that embedding-space proximity is a better predictor of generation

utility than BM25 or cross-encoder score in this setting. On ExpertWritten, Dense accounts for most of the lexical downstream gain; setwise rerankers remain competitive with Dense on token-level metrics (Expert F1 $\approx$ .619–.622, Sim F1 $\approx$ .506), with neither SetRank nor MESA establishing a clear margin. MESA's downstream advantage is most pronounced on multi-facet queries under qualitative evaluation, as the LLM-as-Judge analysis below reports. The primary downstream gain corresponds to the shift from sparse to dense retrieval; set-level modeling contributes an additional but smaller increment, consistent with the improved evidence coverage observed in Tables 1 and 3.

LLM-as-Judge evaluation. Token-F1 and ROUGE-1 measure lexical overlap but are insensitive to *qualitative* differences in answer completeness—particularly for multi-facet queries where covering all relevant aspects matters more than exact wording. To expose these finer-grained quality distinctions, we employ GPT-4o-mini as an LLM judge, scoring each generated answer on three dimensions on a 1–5 Likert scale—**Comprehensiveness**: coverage of all relevant query aspects; **Actionability**: presence of concrete, step-by-step guidance; and **Faithfulness**: absence of hallucinated or unsupported claims. We evaluate the same 200 ExpertWritten queries and additionally stratify by query complexity using the multi-facet subset (N=28, queries with $|\text{Rel}| \geq 2$ gold passages).

Table A13: LLM-as-Judge evaluation (GPT-4o-mini, 1–5 scale) of generated answers from each method's top-5 evidence. Multi-facet queries ($|\text{Rel}| \geq 2$) isolate cases where evidence assembly quality matters most. SetRank is omitted from this evaluation because its generation outputs are derived from the same GPT-4o reader as other methods but SetRank's top-5 sets were not retained after the main experiment; token-level F1/ROUGE for SetRank are reported in Table A12.

	All (N=200)			Multi-facet (N=28)		
Method	**Comp**	**Act**	**Faith**	**Comp**	**Act**	**Faith**
BM25	3.67	3.50	4.89	4.21	4.14	4.96
MonoT5-base	3.67	3.50	4.91	4.36	4.29	4.93
Dense (MiniLM)	4.26	4.08	4.90	4.39	4.11	4.89
MESA (joint)	4.21	4.10	4.89	**4.50**	**4.36**	4.89

Table A13 reveals patterns invisible to token-level metrics. On the full query set, Dense and MESA are indistinguishable on both Comprehensiveness and Actionability, mirroring their near-tied token-level F1 in Table A12. On the multi-facet subset, where evidence assembly is the operative requirement, MESA leads on both judge dimensions, confirming that its coverage advantage translates into more complete answers where it is designed to matter. BM25 and MonoT5 cluster together at substantially lower scores on both axes, consistent with the F1 plateau in Table A12 and reinforcing that cross-encoder reranking alone does not improve evidence assembly quality. These results indicate that F1/ROUGE underestimate MESA's downstream utility on multi-facet queries: the coverage advantage observed in ranking metrics translates into more comprehensive generation precisely where evidence assembly is the operative requirement.

K Encoder Scale-Up: BGE-base-en-v1.5

Table A14: Encoder scale-up results comparing BGE-base-en-v1.5 (109M) and MiniLM (22M) as the frozen backbone. Within each encoder, Dense and MESA differ only in the presence of MESA's set-level module. The same-encoder gap is preserved—and on Simulated, amplified—under the larger encoder. Joint training recipe; mean±std over 3 seeds.

Encoder	**Method**	**nDCG @5 / @10**	**R@5**	**MRR@10**
ExpertWritten (200 queries)				
MiniLM	Dense	0.636 / 0.668	0.802	0.607
(22M)	MESA	**0.741 / 0.765** ±.012 / ±.014	**0.866** ±.012	**0.724** ±.012
	Δ	**+0.105 / +0.097**	**+0.064**	**+0.117**
BGE-base	Dense	0.662 / 0.693	0.814	0.637
(109M)	MESA	**0.765 / 0.788** ±.007 / ±.007	**0.889** ±.009	**0.745** ±.004
	Δ	**+0.103 / +0.095**	**+0.075**	**+0.108**
Simulated (200 queries)				
MiniLM	Dense	0.567 / 0.605	0.758	0.525
(22M)	MESA	**0.642 / 0.671** ±.009 / ±.011	**0.791** ±.010	**0.611** ±.008
	Δ	**+0.075 / +0.066**	**+0.033**	**+0.086**
BGE-base	Dense	0.613 / 0.656	0.755	0.586
(109M)	MESA	**0.732 / 0.758** ±.010 / ±.009	**0.844** ±.004	**0.712** ±.011
	Δ	**+0.119 / +0.102**	**+0.089**	**+0.126**

MESA's set-level advantage is not diminished when the frozen encoder is replaced with BGE-base-en-v1.5 (109M, 768d)—a 5× parameter increase over MiniLM-L6-v2 (22M, 384d). The joint training recipe is repeated with hidden_dim=384 to accommodate the larger input dimensionality; all other details remain identical to the main experiment. The only difference between Dense-BGE and MESA-BGE is the presence of MESA's set-

level module (+5.8M params), which is larger than the base MESA module (≈2.5M) because the increased hidden dimension expands the slot and projection layers; both Dense-BGE and MESA-BGE share the same frozen encoder and identical document embeddings.

MESA's set-level advantage is fully preserved—and on Simulated, amplified—when the base encoder is 5× larger. On ExpertWritten, the absolute gain (ΔnDCG@5=+0.103) is nearly identical to the MiniLM result (+0.105), confirming that set-level reasoning provides orthogonal utility regardless of encoder capacity. On Simulated, the gain is substantially *larger* with BGE (+0.119 vs. +0.075), suggesting that the stronger encoder provides better initial representations that the set-level module can more effectively reorganize for complementary coverage.

These results directly address the concern that a sufficiently strong pointwise encoder might render MESA's coverage mechanism redundant: even with a 109M-parameter encoder that already outperforms the MiniLM baseline on all metrics, Dense retrieval still lacks the cross-document reasoning needed to assemble complementary evidence sets, and the set-level module continues to fill that gap.