THE RICH GET RELEVANT: NEURO-SYMBOLIC DIVERSITY-INDUCING RANKING ON SEMANTIC GRAPHS FOR PERSONALIZED MULTI-DOCUMENT SUMMARIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Query-focused multi-document summarization (MDS) underpins many generative-AI applications (automatic presentations, video and meeting summarization, descriptive question answering, personalized news digests). Yet feeding a long-context LLM raw multi-source text is unreliable: it triggers "lost-in-the-middle" behavior, cross-source redundancy, and hallucination, while classical graph rankers are "feature-blind" and cannot personalize to user intent. We propose RICH, a lightweight ($\sim 10^5$–10^6-parameter) neuro-symbolic ranker that hands the LLM a *structured* ranked list of $(\text{topic}, \text{evidence}, \text{score})$ tuples instead of raw text. RICH extracts intent-filtered atomic topics, organizes them in a hierarchical topic–document graph, and runs a parameter-free Recurrent GNN with intent-aware MLP priors and Graph-Attention transitions that scores each topic by jointly balancing relevance and diversity. It embeds the discrete DivRank random walk inside a differentiable model and offers three simple guarantees: RICH is a differentiable form of Pointwise DivRank; its recurrence is a stable damped walk on the simplex with a score floor and a fixed point (fast convergence measured on all evaluation graphs); and its hierarchical prior gives every source document a positive score floor. Across four DiverseSumm setups and an event-level held-out split with disjoint train/test events, RICH leads strong LLM-prompting and classical content-selection baselines on coverage and diversity, with the largest and statistically significant gains in the query-focused settings. The gains transfer zero-shot to Multi-News and WCEP, and a human study confirms the reliability of the evaluation protocol. RICH adds negligible inference overhead ($\sim$0.3% of end-to-end latency) on top of the LLM calls that every baseline also makes.

1 INTRODUCTION

Query-focused multi-document summarization (MDS) is a foundational primitive behind many recent generative-AI applications: slide-deck and presentation creation (Maheshwari et al., 2024), video and lecture summarization (Liu et al., 2025), podcast summarization (Xiao et al., 2025), long-form question answering over enterprise corpora (Edge et al., 2024; He et al., 2024), and personalized meeting summarization (Kirstein et al., 2024; Chatterjee et al., 2026). These all reduce, at their core, to selecting and synthesizing the most relevant content from many heterogeneous sources in response to a user query (or intent). In real products the query is not merely a retrieval filter but the natural *personalization* interface, letting the same corpus be summarized differently for different users. Solving this well requires jointly satisfying three competing objectives: comprehensive *coverage* of the source material, semantic *diversity* across surfaced viewpoints, and strict *personalization* to the user intent.

Despite their fluency, state-of-the-art LLMs degrade sharply on long, multi-document inputs fed as a flat concatenation: the "lost-in-the-middle" phenomenon (Liu et al., 2024; Hsieh et al., 2024) makes them ignore evidence away from the prompt boundaries, and GPT-4 covers under 40% of the diverse information in multi-document news collections (Huang et al., 2024). These failures are not merely a context-length artefact; they persist within long-context windows and worsen for the

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.

smaller, deployable LMs that products favour, where attention diffuses across redundant passages and triggers hallucinations (Laban et al., 2022; Goyal et al., 2022). Classical extractive and graph-based methods rank salient segments *before* generation (Erkan & Radev, 2004; Carbonell & Goldstein, 1998; Mei et al., 2010), and vertex-reinforced random walks (Pemantle, 1992; 2007) even guarantee diversity through a "rich-gets-richer" competition. But they are *feature-blind* (operating on discrete topology, discarding dense semantics, and unable to condition on a continuous intent vector) and largely *non-differentiable* (their heuristic kernels admit no end-to-end gradient, so they cannot learn from downstream feedback). As a result, neither pure neural generation nor purely symbolic ranking delivers the *coverage*, *diversity*, and intent-relevance that MDS demands.

We propose RICH, a lightweight ($\sim 10^5$–10^6-parameter) neuro-symbolic ranker that supplies the LLM with a *structured* shortlist of atomic topics, their evidence spans, and importance scores. Source documents are decomposed into intent-filtered atomic claims arranged in a hierarchical semantic graph that couples a topic graph G_t with a document graph G_a via a parent map. Each level carries an intent-aware MLP prior and a Graph Attention kernel (Veličković et al., 2018); a parameter-free Recurrent GNN (Scarselli et al., 2009; Li et al., 2016) unrolls a coupled, history-aware message passing that absorbs redundant cliques while pulling mass toward intent-relevant topics, trained with a Determinantal Point Process log-likelihood objective (Kulesza & Taskar, 2012). The key design principle is to embed the *discrete*, non-differentiable DivRank walk (Mei et al., 2010) inside a differentiable model: unrolling it as a finite recurrent computation makes it end-to-end trainable, so gradients from custom self-supervised objectives reshape the walk while its stochastic, diversity-enforcing dynamics are preserved exactly.

Contributions. (1) We introduce RICH (**R**ecurrent **I**ntent-**C**onditioned **H**eterogeneous ranking), a lightweight neuro-symbolic ranker for query-focused MDS that mitigates lost-in-the-middle degradation by feeding the LLM a structured, evidence-grounded ranked shortlist instead of raw multi-document text; as a backbone-agnostic pre-ranker, RICH works with any generation model, open-source (e.g. Llama) or proprietary (e.g. GPT-4o), without modification. **(2)** We make the discrete, non-differentiable DivRank random walk *differentiable* and train it end-to-end: a parameter-free, hierarchical Recurrent GNN unrolls the walk so that gradients flow through it via backpropagation-through-time, learning the topic- and document-level graph geometry while preserving the walk's stochastic, diversity-enforcing dynamics exactly. Three custom self-supervised objectives (an intent-coverage loss, a determinantal-point-process set-level diversity loss, and a sparsity regularizer) then enforce the desired ranking properties without any labels. **(3)** We give three simple, self-contained guarantees (§5): RICH is a differentiable form of Pointwise DivRank (Mei et al., 2010); its recurrence is a stable damped walk on the simplex with a guaranteed score floor and a fixed point, with fast convergence measured directly on all evaluation graphs; and its hierarchical prior gives every source document a positive score floor. We keep the statements deliberately simple and prove them in full. **(4)** Across four evaluation setups on DiverseSumm (Huang et al., 2024), an event-level held-out split with disjoint train/test events, and zero-shot transfer to Multi-News and WCEP, RICH leads standalone LLM prompting and classical content-selection baselines on coverage and diversity, with statistically significant gains in the query-focused settings and negligible inference overhead. A human study validates the evaluation protocol.

2 Related Work

Graph-based content selection. Classical rankers such as LexRank (Erkan & Radev, 2004) and TextRank (Mihalcea & Tarau, 2004), and diversity-aware rerankers such as MMR (Carbonell & Goldstein, 1998) and DivRank (Mei et al., 2010), rank text units by graph centrality. They treat the graph as flat and fixed, operate on discrete topology, and admit no gradient signal. RICH differs in two ways: it couples a topic-level graph with a document-level graph, and it makes prestige a *learned, recurrent* quantity rather than a fixed PageRank-style value, so the routing geometry is optimized end-to-end.

Determinantal point processes (DPPs). DPP selection (Kulesza & Taskar, 2012) and its recent use for multi-document key-point selection (Padmakumar et al., 2025) trade off coverage and diversity with a kernel, and DPP objectives have also been used to post-train LLMs for diverse, high-quality generation (Chen et al., 2026b). Our DPP term plays a similar role, but it sits *inside* the learned

recurrent updates as a training objective rather than as a one-shot selection step applied after ranking; an ablation isolates its effect (§6.2).

Recurrent GNNs and reinforced walks. We frame message passing as a Recurrent GNN (Scarselli et al., 2009; Li et al., 2016) that unrolls a vertex-reinforced random walk (Pemantle, 1992; 2007). Prior GNN rerankers use a single flat graph; RICH adds a hierarchical, parameter-light recurrent coupling between topics and documents with convergence guarantees (§5).

Long-context and RAG. Our motivation connects to long-context "lost-in-the-middle" analyses (Liu et al., 2024; Hsieh et al., 2024) and to benchmarks that isolate a model's genuine long-context capacity from its stored knowledge (Chen et al., 2026c), as well as to context ranking for retrieval-augmented generation (Yu et al., 2024; Edge et al., 2024; He et al., 2024), including recent analyses of when graph structure actually helps RAG (Xiang et al., 2026) and multi-granularity retrieval for long-context summarization (Chen et al., 2026a). By ranking and compressing useful, diverse content *before* generation, RICH reduces the position and length problems that a downstream LLM would otherwise face.

3 Problem Statement and Notation

Let $\mathcal{D} = \{d_1, \ldots, d_K\}$ be a collection of K source documents on a shared theme, and let I denote a user intent encoding the personalized information need. To bypass the lost-in-the-middle degradation that LLMs suffer on the raw concatenation of $\mathcal{D}$, we cast generation as a *discrete subset-selection and routing* problem on the universe $\mathcal{V} = \{v_1, \ldots, v_n\}$ of atomic topics (indivisible factual claims) extracted from $\mathcal{D}$, with parent map $a : \mathcal{V} \to \mathcal{A}$ recording the source document of each topic. Our goal is to learn a ranking function $\pi : \mathcal{V} \to [0, 1]$ such that the resulting ranked list of $(v, \text{evidence}(v), \pi(v))$ tuples jointly maximizes two competing objectives: **(i) intent relevance**, high $\pi(v)$ implies semantic alignment with I; and **(ii) semantic diversity**, the head of the ranking is non-redundant and represents the differing viewpoints in $\mathcal{D}$. This ranked list is then handed to a downstream LLM as *structured* input in lieu of raw $\mathcal{D}$, so the LLM is freed from implicit content selection or diversity penalization. Because optimizing relevance and diversity simultaneously over $\mathcal{V}$ is combinatorially intractable, we formulate π as the stationary distribution of a diversity-inducing, time-variant random walk over a semantic knowledge graph (§4). A full symbol glossary is in Appendix E.

4 Methodology: Neuro-Symbolic Diversity-Inducing Graph Ranking

Our approach solves the problem through a progressive, six-stage pipeline (Figure 1). We leverage neural networks to map the continuous semantic space of text and intent, while employing symbolic mathematical rules to govern the discrete, diversity-enforcing routing logic.

4.1 Intent-Filtered Atomic Topic Extraction and Hierarchical Graph

Topic extraction (S1). We process each document $d_i \in \mathcal{D}$ independently using an LLM to extract a set of atomic key points. To filter noise and hallucinations, we embed each extracted key point into a dense vector $x_v \in \mathbb{R}^d$ and the user intent into $x_I \in \mathbb{R}^d$ using a pre-trained dense retriever, compute their cosine similarity, and apply a threshold τ_{extract} to discard irrelevant topics, ensuring the initial node set is grounded in the user's intent.

Hierarchical graph (S2). A flat topic-level graph encourages diversity but offers no structural guarantee that the head of the ranking spans *multiple* source documents; if several high-centrality topics come from one dominant document, the LLM rewriter over-allocates attention to it and undercovers the rest. We therefore couple two graphs. The *topic graph* $G_t = (\mathcal{V}, E_t)$ takes atomic topics as nodes and forms an undirected edge (u, v) when $\cos(x_u, x_v) \geq \tau_{\text{graph}}$ or via k-nearest neighbours, defining the neighbourhood $\mathcal{N}(v)$. The *document graph* $G_a = (\mathcal{A}, E_a)$ over source documents $\mathcal{A} = \{a_1, \ldots, a_K\}$ uses the topic-centroid embedding $\mathbf{x}_{a_i} = |\mathcal{V}_i|^{-1} \sum_{v \in \mathcal{V}_i} x_v$ with $\mathcal{V}_i = \{v : a(v) = a_i\}$; edges connect pairs with $\cos(\mathbf{x}_{a_i}, \mathbf{x}_{a_j}) \geq \tau_a$, augmented by k-NN for sparser events. The parent map $a(\cdot)$ provides a deterministic document→topic pathway through which document prestige is later injected into the topic-level prior (§4.3).

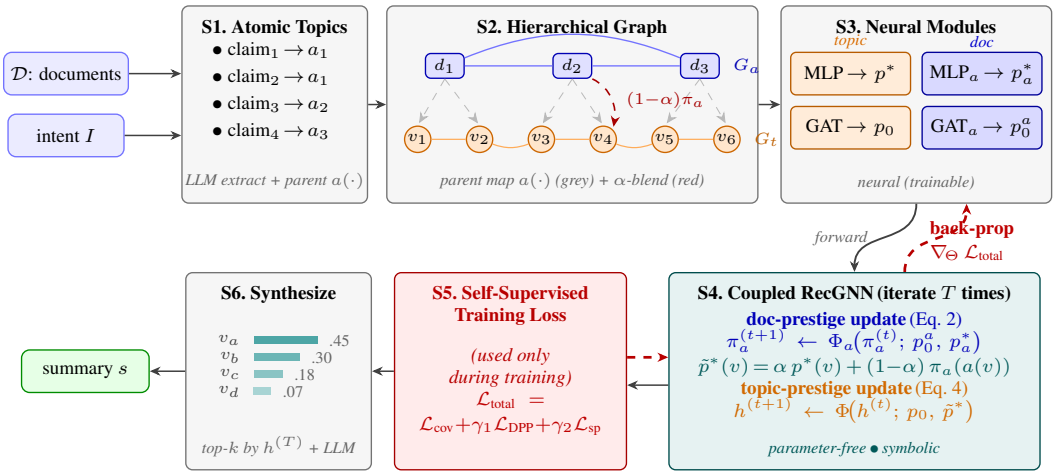

Figure 1: **The RICH pipeline. S1** extracts atomic claims and links each to its source document. **S2** couples a topic-level graph (orange) with a document-level graph (blue) so document-level prestige can be injected into the topic prior. **S3** provides the trainable neural components: intent-aware priors and semantic transition kernels at both levels. **S4** runs a parameter-free coupled recurrent solver that simultaneously enforces relevance and diversity. **S5 (red, training-only)** computes the self-supervised loss and back-propagates gradients to the trainable modules; it is bypassed at inference. **S6** hands the top-ranked topics to an LLM that composes the final summary.

4.2 Intent-Aware Priors and Semantic Transition Kernels

Priors (relevance engine, S3a). We compute intent-conditioned priors at *both* levels of the hierarchy that act as "teleportation" anchors pulling the random walks back toward the user query. A small MLP yields the topic prior $p^*(v) = \mathrm{softmax}(\mathrm{MLP}([x_v \,\|\, x_I]))$ ($\|$ denotes concatenation), and a parameter-disjoint MLP_a produces the document prior $p^*_a(a_i) = \mathrm{softmax}_i(\mathrm{MLP}_a([\mathbf{x}_{a_i} \,\|\, x_I]))$. When no intent is provided (generic summarization), we substitute PageRank (Page et al., 1999) centralities of G_t and G_a, which favour semantically central, broadly shared, high-coverage content.

Transitions (structure engine, S3b). We replace the degree-based structural heuristic of classical walks with a learned semantic flow at each level. A Graph Attention Network (Veličković et al., 2018) over G_t computes the topic-level transition kernel:

$$e_{uv} = \mathrm{LeakyReLU}\big(\mathbf{a}^\top [\mathbf{W}x_u \,\|\, \mathbf{W}x_v]\big), \qquad p_0(u,v) = \frac{\exp(e_{uv})}{\sum_{k\in\mathcal{N}(u)} \exp(e_{uk})}, \tag{1}$$

where $\mathbf{W}$ is a learnable weight matrix and $\mathbf{a}$ a learnable attention vector. A second GAT (disjoint parameters $\mathbf{W}_a, \mathbf{a}_a$) over G_a computes the document-level kernel $p^a_0(a_i, a_j)$. Both kernels are row-stochastic.

4.3 Diversity through a Coupled Recurrent GNN

Standard GNNs pass messages with static weights and cannot penalize semantic redundancy. Following Scarselli et al. (2009); Li et al. (2016), we frame message passing as a Recurrent GNN that applies the same parameters iteratively until node states reach a stationary distribution. We unroll a *vertex-reinforced* random walk so a high-prestige topic absorbs the probability mass of its semantically similar neighbours, enforcing a "rich-gets-richer" competition that penalizes redundant cliques. To additionally guarantee multi-document coverage at the head of the ranking, we run two such walks *jointly*, one over G_a and one over G_t, coupled via the parent map $a(\cdot)$.

Document-level prestige update (S4). Initialize $\pi^{(0)}_a = p^*_a$. The document prestige evolves under the DivRank recurrence:

$$\pi^{(t+1)}_a(a_i) \;=\; (1-\lambda_a)\, p^*_a(a_i) \;+\; \lambda_a \sum_{a_j \in \mathcal{N}(a_i)} \pi^{(t)}_a(a_j) \cdot \frac{p^a_0(a_j, a_i)\, \pi^{(t)}_a(a_i)}{\sum_{a_\ell \in \mathcal{N}(a_j)} p^a_0(a_j, a_\ell)\, \pi^{(t)}_a(a_\ell)}, \tag{2}$$

where $\lambda_a \in (0,1)$ is the document-level damping factor. With $K \leq 10$ documents per event, this adds negligible compute.

Document-augmented topic prior. The topic baseline blends the intent prior with the running document prestige of the parent, spread over the parent's topics:

$$\tilde{p}^*(v) \;=\; \alpha\, p^*(v) \;+\; (1{-}\alpha)\,\frac{\pi_a^{(t)}\big(a(v)\big)}{|\mathcal{V}_{a(v)}|}, \qquad \alpha \in [0,1]. \tag{3}$$

Dividing the parent prestige by its number of topics $|\mathcal{V}_{a(v)}|$ keeps $\tilde{p}^*$ a probability distribution ($\sum_v \tilde{p}^*(v) = 1$), so the topic update below is an exact Pointwise DivRank step (Theorem 1). Setting $\alpha = 1$ recovers a flat formulation; a non-trivial α lifts every topic's floor prestige by a share of its parent document's prestige, so no source document can be starved in the ranking (Theorem 3).

Topic-level prestige update. Let $h_v^{(t)}$ denote the topic prestige at step t, with $h_v^{(0)} = \tilde{p}^{*\,(0)}(v)$. The message from u to v is dynamically scaled by v's current prestige, and the update blends the incoming aggregate with the document-augmented prior under damping $\lambda \in (0,1)$:

$$m_{u\to v}^{(t)} \;=\; \frac{p_0(u,v)\,h_v^{(t)}}{\sum_{k\in\mathcal{N}(u)} p_0(u,k)\,h_k^{(t)}}\,h_u^{(t)}, \qquad h_v^{(t+1)} \;=\; (1{-}\lambda)\,\tilde{p}^*(v) \;+\; \lambda \sum_{u\in\mathcal{N}(v)} m_{u\to v}^{(t)}. \tag{4}$$

The two updates run jointly for T steps, with $\tilde{p}^*$ refreshed at each step using the current $\pi_a^{(t)}$. The final state $h_v^{(T)}$ is our ranking function $\pi(v)$. Crucially, this stage introduces *no new parameters*: the network's expressive power is entirely channelled into the semantic geometry (p_0, p_0^a, p^*, p_a^*), while message passing strictly obeys the symbolic, diversity-enforcing constraints of the vertex-reinforced walk. The hierarchical fixed point can be viewed as the joint stationary distribution of the two coupled DivRank walks (Appendix B.3), which exists and is stable by Theorem 2. Training (Alg. 1) and inference (Alg. 2) are in Appendix A.

4.4 Structured Hand-off and Self-Supervised Training

Hand-off (S6). We sort all topics by descending $h_v^{(T)}$ and emit, for each v, a tuple $\big(v,\ \text{evidence}(v),\ h_v^{(T)}\big)$ comprising the topic statement, its supporting evidence span(s), and its importance score (already balancing relevance and diversity). The ranked list is passed to a downstream LLM that composes the summary s. By providing structured, ranked, evidence-grounded input rather than raw documents, the LLM is freed from implicit content selection and can focus on faithful prose. RICH comprises only two MLPs and two GAT layers, on the order of 10^5–10^6 trainable parameters, at least three orders of magnitude smaller than the dense retriever and the LLM rewriter; end-to-end inference is dominated by a single LLM synthesis call (§6.2, Appendix G).

Training objectives (S5). Since atomic-topic rankings are not annotated, we train the GAT/MLP modules end-to-end on a self-supervised loss over the converged prestige $h^{(T)}$, combining three custom objectives that each enforce one desired property of the ranking (intent coverage, set-level diversity, and non-degeneracy). Define the *predicted summary embedding* $z_{\text{sum}} = \sum_v h_v^{(T)} x_v$ and the *intent-targeted embedding* $z_{\text{target}} = \sum_v w_v x_v / \sum_v w_v$ with $w_v = \max(0, \cos(x_v, x_I))$. *(i) Intent coverage:* $\mathcal{L}_{\text{cov}} = 1 - \cos(z_{\text{sum}}, z_{\text{target}})$. *(ii) DPP log-likelihood diversity:* following Kulesza & Taskar (2012), we form the symmetrised GAT kernel $K_\Theta[u,v] = \frac{1}{2}(p_0(u,v){+}p_0(v,u))$ and the quality-diversity L-ensemble $L_\Theta = \text{diag}(\sqrt{h^{(T)}})\,K_\Theta\,\text{diag}(\sqrt{h^{(T)}})$, and minimize

$$\mathcal{L}_{\text{DPP}} = -\log\det\big(L_\Theta + \varepsilon I\big), \tag{5}$$

with ridge $\varepsilon{=}10^{-5}$. A first-order expansion gives $\mathcal{L}_{\text{DPP}} \approx -\sum_v \log h_v^{(T)} + \sum_{u\neq v} h_u^{(T)} h_v^{(T)} K_\Theta[u,v]$: the second term is the classical pairwise diversity penalty, the first an additional log-coverage regularizer (Appendix C). *(iii) Sparsity:* a normalized negative-entropy term $\mathcal{L}_{\text{sparsity}} = -\frac{1}{\log|\mathcal{V}|}\sum_v h_v^{(T)} \log(h_v^{(T)} + \epsilon)$ rewards a peaky ranking and breaks the uniform-mode symmetry. The total objective is $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cov}} + \gamma_1 \mathcal{L}_{\text{DPP}} + \gamma_2 \mathcal{L}_{\text{sparsity}}$. The Markov property keeps $\sum_v h_v^{(T)} = 1$, so all terms are intrinsically bounded (Appendix D).

5 THEORETICAL ANALYSIS

RICH rests on three simple, self-contained guarantees. We state them here and prove them in full in Appendix B. Throughout, the topic and document graphs are connected and undirected (edges symmetrised, self-loops on isolated nodes), the damping factors satisfy $\lambda, \lambda_a \in (0, 1)$, and the priors p^*, p_a^* are strictly positive softmax outputs; the hierarchical prior $\tilde{p}^*$ (Eq. 3) is a probability distribution by construction.

(T1) RICH is a differentiable Pointwise DivRank. The core of RICH is the classical DivRank walk (Mei et al., 2010) with its two hand-set ingredients, the transition kernel and the prior, replaced by learned modules. Unrolling the walk for a finite number of steps turns a discrete, non-differentiable random walk into an end-to-end trainable layer.

Theorem 1 (RICH is Pointwise DivRank made differentiable)**.** *Fix the parameters* $\Theta = \{$*GAT, MLP, GAT*$_a$*, MLP*$_a\}$*. Then the* RICH *topic update (Eq. 4) is exactly the Pointwise DivRank recurrence of Mei et al. (2010, Eq. 10) with transition kernel* p_0 *and prior* $\tilde{p}^*$*, the only change being that* p_0 *(a GAT) and* $\tilde{p}^*$ *(an intent MLP blended with the document walk) are differentiable functions of* Θ*. Consequently the* T*-step map* $h^{(T)}(\Theta)$ *is differentiable in* Θ *almost everywhere, and any downstream loss* $\mathcal{L}_{\text{total}}(h^{(T)})$ *is trained by backpropagation through time (proof in Appendix B.1).*

(T2) The ranking is stable and has a fixed point. Each step is a damped average: a fraction $1-\lambda$ of a strictly positive prior plus a fraction λ of a mass-preserving walk step. This keeps every score a valid distribution, keeps it above a positive floor, and guarantees a stationary ranking. We do not claim a closed-form contraction rate; we measure it directly instead.

Theorem 2 (Existence and stability of the RICH recurrence)**.** *Under the assumptions above, the coupled document/topic update* $\Phi(\pi, h) = (\Psi(\pi), \Phi_h(h; \pi))$ *satisfies:* ***(i)*** *it maps the product of probability simplices into itself;* ***(ii)*** *every coordinate stays above a positive floor,* $h_v^{(t)} \geq (1-\lambda)\,\alpha \min_v p^*(v) =: \delta > 0$ *and* $\pi_{a_i}^{(t)} \geq (1-\lambda_a) \min_i p_a^*(a_i) =: \delta_a > 0$*, for all* t*; and* ***(iii)*** Φ *has a fixed point* (π^*, h^*)*, and because the floor keeps every denominator positive, the unrolled solver and its BPTT gradients are well defined (proof in Appendix B.2).*

Theorem 2 gives existence and stability, not a rate. We therefore measure the rate directly: averaged over all 245 evaluation graphs, the iterates approach the fixed point geometrically at rate $\rho \approx 0.73$, and by the default $T{=}10$ the per-step change has fallen to ≈ 0.008 in total variation (Appendix B.2, Table 5).

(T3) The hierarchy gives every source a floor. The document-augmented prior (Eq. 3) injects a share of each document's prestige into its topics, and the damping re-adds it at every step, so every source keeps a positive floor that topic-level competition cannot erase.

Theorem 3 (Per-source prestige floor)**.** *Under the hierarchical prior with* $\alpha < 1$*, every topic* v *keeps, at every step,*

$$h_v^{(t+1)} \;\geq\; (1-\lambda)\,(1-\alpha)\,\frac{\pi_a^{(t)}(a(v))}{|\mathcal{V}_{a(v)}|} \;>\; 0.$$

Hence each source document a_i *guarantees each of its topics a positive score floor equal to its per-topic share of the document prestige, so no source can be driven to zero by topic-level competition. Setting* $\alpha = 1$ *removes this floor (proof in Appendix B.4).*

Theorem 3 is a per-source floor, not a coverage-dominance claim. The broad multi-document coverage RICH achieves in practice comes mainly from the *diversity* (DPP) term rather than the α-blend alone: removing $\mathcal{L}_{\text{DPP}}$ drops top-k document coverage by 0.09 at $k{=}10$ ($p{=}0.011$), 0.20 at $k{=}15$, and 0.26 at $k{=}20$ (both $p{<}0.001$), whereas the α-blend alone shows no significant per-sample gap on the small ablation set (Appendix J). The two failure cases in Appendix O are where one dense document pair dominates despite the floor, so they lie outside this guarantee rather than contradict it.

6 EXPERIMENTS

Our experimental framework isolates the content-selection phase from generation, comparing our neuro-symbolic ranker against both purely generative models and classical discrete selection under identical rewriting prompts and judges.

6.1 Datasets, Setups, Metrics, and Baselines

Data. We evaluate on **DiverseSumm** (Huang et al., 2024): 245 news events, each with 10 source articles and 1–10 human-authored questions (611 total). Since 78.3% of events have ≤ 3 questions (insufficient for reliable coverage estimation), we augment each event with 10 GPT-5.1-generated factual Q&A pairs per article ($\approx$100 per event, $\approx$24,500 total), forming **DiverseSumm Augmented**. For query-focused evaluation, we generate 5 synthetic user intents per event and retain those with at least 20 intent-relevant augmented questions (cosine ≥ 0.6), producing 480 intents across 245 events. A human study (Appendix I) validates these synthetic artefacts.

Four setups span two dimensions: summarization mode (generic vs. query-focused) and question set (original vs. augmented). **S1**: generic summary, original sparse questions (reproducing the protocol of Padmakumar et al., 2025). **S2**: generic summary, all augmented questions (higher statistical power). **S3**: query-focused summary evaluated against *all* augmented questions (does personalization hurt breadth?). **S4**: query-focused summary evaluated against only the *intent-relevant* questions (strictest personalization test). To prevent leakage, RICH is trained on generic samples and synthetic intents only; the evaluation Q&A pairs are never seen during training.

Metrics. **Topic Cluster Coverage (TCC)**: source key points are embedded with `all-MiniLM-L6-v2` and grouped into 10 K-Means clusters; a cluster is covered if any summary sentence has cosine ≥ 0.5 with its centroid. **Semantic Spread (SS)**: mean cosine distance of summary-sentence embeddings from their centroid. Both are computed directly on the summary without an LLM judge and are stable across backbones. **Coverage** and **Answerability** are LLM-judged by GPT-5.1 (Padmakumar et al., 2025): for each question q with reference answer a, whether q is answerable from s (Answerability) and whether the answer is factually consistent with a (Coverage). Coverage is our primary metric; Answerability is a secondary informativeness diagnostic.

Baselines. We benchmark against three baselines from Padmakumar et al. (2025) using the same rewriting prompt and GPT-5.1 judge, with Llama-3.1-8B and GPT-4o backbones: **Naive LLM** (all articles concatenated, implicit selection); **LLM-KPs** (LLM extracts one-sentence key points, then a second LLM call freely selects any subset); and **LLM+DPP** (Padmakumar et al., 2025) (a DPP selects a diverse subset in embedding space, with intent relevance via L-ensemble reweighting in the query-focused variant). All methods, including RICH, feed their selected content to the *same* final rewriting call, so differences reflect content-selection quality alone.

Implementation. RICH uses `intfloat/e5-large-v2` (Apache-2.0, d=1024) to embed topics and intents. The RecGNN is unrolled for T=10 with damping λ=λ_a=0.85 and prior blend α=0.6; loss weights are γ_1=0.5, γ_2=0.1. Within each results table a single model serves as both the topic extractor and the final generator: Table 1 uses Llama-3.1-8B and Table 2 uses GPT-4o. Full hyperparameters are in Appendix D.

6.2 Main Results, Generalization, and Analysis

Main results (Table 1). With the Llama-3.1-8B backbone, RICH leads on all four metrics in every setup, with absolute Coverage gains of 0.05–0.06 and Answerability gains of 0.05–0.11 over LLM+DPP, the strongest baseline on Coverage. With a GPT-4o backbone (Table 2), RICH retains the Coverage lead (0.01–0.04) and the best Semantic Spread; LLM+DPP edges RICH on Answerability by ≤ 0.01. The substantially larger margins under the smaller backbone reflect a structural bottleneck: an 8B-parameter LLM has limited capacity to distil relevant, coherent content from raw articles, so RICH's structured, diversity-aware pre-selection yields a greater marginal benefit. A more capable backbone partly compensates, narrowing but not eliminating the Coverage advantage. We next test whether these gains hold on *unseen* events (Table 3).

Generalization to unseen events (Table 3). Because RICH learns a ranker, a key question is whether its gains hold on events unseen in training rather than reflecting adaptation to the training events. We construct a fixed event-level split (196 train / 49 held-out, disjoint), retrain RICH on the 196 training events, and evaluate on the 49 held-out events; baselines never train on events. On these unseen events, RICH leads LLM+DPP on Coverage in *all four* setups, and the lead is *statistically significant in both query-focused settings* (S3: $+0.072$, p=0.001; S4: $+0.057$, p=0.014), precisely where personalization matters. The held-out-trained model reproduces the reported (245-event) model on the same 49 events to within $\approx$ 0.00–0.06 Coverage per setup, evidence of generalization

Table 1: Results with the **Llama-3.1-8B** backbone (245 events; S1–S2 report SE over events, S3–S4 over 480 intent–event pairs). TCC/SS are non-LLM metrics; Coverage/Answerability are judged by GPT-5.1. **Bold** = best per metric per setup; underline = second best.

Setup	Method	TCC	SS	Coverage	Answerability
S1: Generic, Original Q&A	Naive LLM	$0.760_{\pm.009}$	$0.339_{\pm.004}$	$0.248_{\pm.022}$	$0.307_{\pm.024}$
	LLM-KPs	$0.791_{\pm.008}$	$0.354_{\pm.003}$	$0.572_{\pm.024}$	$\underline{0.658}_{\pm.022}$
	LLM+DPP	$\underline{0.797}_{\pm.007}$	$\underline{0.364}_{\pm.003}$	$\underline{0.574}_{\pm.024}$	$0.646_{\pm.023}$
	RICH	$\mathbf{0.829}_{\pm.008}$	$\mathbf{0.377}_{\pm.004}$	$\mathbf{0.627}_{\pm.024}$	$\mathbf{0.754}_{\pm.020}$
S2: Generic, Augmented Q&A	Naive LLM	$0.760_{\pm.009}$	$0.339_{\pm.004}$	$0.217_{\pm.009}$	$0.262_{\pm.010}$
	LLM-KPs	$0.791_{\pm.008}$	$0.354_{\pm.003}$	$0.443_{\pm.011}$	$0.512_{\pm.011}$
	LLM+DPP	$\underline{0.797}_{\pm.007}$	$\underline{0.364}_{\pm.003}$	$\underline{0.462}_{\pm.011}$	$\underline{0.547}_{\pm.012}$
	RICH	$\mathbf{0.829}_{\pm.008}$	$\mathbf{0.377}_{\pm.004}$	$\mathbf{0.519}_{\pm.013}$	$\mathbf{0.613}_{\pm.012}$
S3: Query-Focused, All Aug. Q&A	Naive LLM	$0.750_{\pm.009}$	$0.333_{\pm.003}$	$0.331_{\pm.011}$	$0.397_{\pm.012}$
	LLM-KPs	$0.761_{\pm.008}$	$0.329_{\pm.003}$	$0.370_{\pm.012}$	$0.447_{\pm.013}$
	LLM+DPP	$\underline{0.812}_{\pm.008}$	$\underline{0.361}_{\pm.003}$	$\underline{0.432}_{\pm.011}$	$\underline{0.543}_{\pm.012}$
	RICH	$\mathbf{0.817}_{\pm.008}$	$\mathbf{0.370}_{\pm.003}$	$\mathbf{0.495}_{\pm.013}$	$\mathbf{0.594}_{\pm.013}$
S4: Query-Focused, Relevant Q&A	Naive LLM	$0.750_{\pm.009}$	$0.333_{\pm.003}$	$0.389_{\pm.012}$	$0.452_{\pm.012}$
	LLM-KPs	$0.761_{\pm.008}$	$0.329_{\pm.003}$	$0.392_{\pm.012}$	$0.478_{\pm.013}$
	LLM+DPP	$\underline{0.812}_{\pm.008}$	$\underline{0.361}_{\pm.003}$	$\underline{0.446}_{\pm.011}$	$0.570_{\pm.011}$
	RICH	$\mathbf{0.817}_{\pm.008}$	$\mathbf{0.370}_{\pm.003}$	$\mathbf{0.509}_{\pm.013}$	$\mathbf{0.629}_{\pm.013}$

Table 2: Results with the **GPT-4o** backbone (245 events). Coverage and Answerability judged by GPT-5.1. **Bold** = best per metric per setup; underline = second best.

Setup	Method	TCC	SS	Coverage	Answerability
S1: Generic, Original Q&A	Naive LLM	$0.817_{\pm.008}$	$0.402_{\pm.003}$	$0.664_{\pm.023}$	$0.827_{\pm.018}$
	LLM-KPs	$0.811_{\pm.008}$	$0.395_{\pm.003}$	$0.742_{\pm.021}$	$0.841_{\pm.018}$
	LLM+DPP	$\underline{0.847}_{\pm.008}$	$\underline{0.425}_{\pm.003}$	$\underline{0.802}_{\pm.019}$	$\mathbf{0.900}_{\pm.014}$
	RICH	$\mathbf{0.849}_{\pm.007}$	$\mathbf{0.429}_{\pm.002}$	$\mathbf{0.814}_{\pm.018}$	$\underline{0.892}_{\pm.015}$
S2: Generic, Augmented Q&A	Naive LLM	$0.817_{\pm.008}$	$0.402_{\pm.003}$	$0.617_{\pm.012}$	$0.850_{\pm.010}$
	LLM-KPs	$0.811_{\pm.008}$	$0.395_{\pm.003}$	$0.641_{\pm.012}$	$0.857_{\pm.010}$
	LLM+DPP	$\underline{0.847}_{\pm.008}$	$\underline{0.425}_{\pm.003}$	$\underline{0.773}_{\pm.011}$	$\mathbf{0.928}_{\pm.006}$
	RICH	$\mathbf{0.849}_{\pm.007}$	$\mathbf{0.429}_{\pm.002}$	$\mathbf{0.788}_{\pm.010}$	$\underline{0.924}_{\pm.007}$
S3: Query-Focused, All Aug. Q&A	Naive LLM	$0.793_{\pm.012}$	$0.384_{\pm.003}$	$0.564_{\pm.016}$	$0.798_{\pm.013}$
	LLM-KPs	$0.745_{\pm.011}$	$0.366_{\pm.004}$	$0.471_{\pm.016}$	$0.684_{\pm.016}$
	LLM+DPP	$\mathbf{0.839}_{\pm.010}$	$\underline{0.414}_{\pm.004}$	$\underline{0.703}_{\pm.012}$	$\mathbf{0.906}_{\pm.007}$
	RICH	$\underline{0.836}_{\pm.009}$	$\mathbf{0.418}_{\pm.003}$	$\mathbf{0.746}_{\pm.012}$	$\underline{0.898}_{\pm.008}$
S4: Query-Focused, Relevant Q&A	Naive LLM	$0.793_{\pm.012}$	$0.384_{\pm.003}$	$0.582_{\pm.015}$	$0.812_{\pm.013}$
	LLM-KPs	$0.745_{\pm.011}$	$0.366_{\pm.004}$	$0.485_{\pm.016}$	$0.709_{\pm.015}$
	LLM+DPP	$\mathbf{0.839}_{\pm.010}$	$\underline{0.414}_{\pm.004}$	$\underline{0.718}_{\pm.012}$	$\mathbf{0.915}_{\pm.007}$
	RICH	$\underline{0.836}_{\pm.009}$	$\mathbf{0.418}_{\pm.003}$	$\mathbf{0.759}_{\pm.012}$	$\underline{0.908}_{\pm.008}$

rather than memorization. The only cell where RICH trails is S1 Coverage (vs. LLM-KPs), the unconstrained-budget generic regime where all selection methods feed a similar number of key points to the rewriter. This event-level protocol is more demanding and practically relevant than holding out only question–answer pairs; the full study (protocol, GPT-4o results, reproduction reference) is in Appendix H.

Zero-shot transfer. Applied *without fine-tuning* to Multi-News (Fabbri et al., 2019) and WCEP (Gholipour Ghalandari et al., 2020) (25 events each, Setup-2 protocol), the DiverseSumm-trained RICH leads on all four metrics on both corpora (Table 4): Coverage 0.686 on Multi-News and 0.602 on WCEP, ahead of the best baseline by 0.05–0.06. Gains over LLM+DPP are significant ($p<0.001$ on Multi-News; $p=0.001/0.010$ on WCEP); per-comparison p-values are in Appendix F.

Ablations. On a 30-event subset, removing the DPP loss consistently reduces Coverage, collapsing the RecGNN to $T=1$ lowers Coverage and Answerability in every setup (and TCC/SS in the generic setups), and removing the hierarchical prior causes the largest generic-Coverage drop (S1: -0.14). At $n=30$ most pairwise differences are not individually significant, so we report the ablation as a directional trend; the primary component evidence comes from the larger-scale comparisons (Appendix K).

Evaluation reliability (human study). Three independent expert annotators, blind to the judge, validated the synthetic artefacts and the GPT-5.1 judge on a fixed stratified sample: 95.6% of synthetic questions were accepted as answerable and on-topic, and human–judge agreement was 82%

Table 3: Event-level **held-out** generalization (Llama-3.1-8B backbone, GPT-5.1 judge): RICH is trained on 196 events and evaluated on **49 disjoint held-out events**; baselines never train on events. Setups **S1/S2**: generic summaries on original/augmented Q&A; **S3/S4**: query-focused summaries on all/intent-relevant Q&A (Cov. = Coverage, Ans. = Answerability). */*** mark RICH significantly above LLM+DPP on Coverage ($p<.05$/.001; paired bootstrap, 10k). **Bold** = best per metric per setup.

	S1		S2		S3		S4	
Method	Cov.	Ans.	Cov.	Ans.	Cov.	Ans.	Cov.	Ans.
Naive LLM	0.249	0.270	0.227	0.263	0.302	0.361	0.355	0.403
LLM-KPs	**0.572**	**0.650**	0.410	0.470	0.333	0.413	0.337	0.420
LLM+DPP	0.494	0.595	0.438	0.526	0.372	0.491	0.403	0.529
RICH	0.534	**0.650**	**0.456**	**0.542**	**0.444*****	**0.511**	**0.460***	**0.540**

Table 4: Zero-shot out-of-distribution generalization on Multi-News and WCEP (25 events each, Llama-3.1-8B backbone, GPT-5.1 judge, Setup-2 protocol). RICH uses the DiverseSumm-trained checkpoint with no fine-tuning. **Bold** = best per metric per dataset; underline = second best. Per-comparison significance tests are in Appendix F.

Dataset	Method	TCC	SS	Coverage	Answerability
Multi-News	Naive LLM	$0.584_{\pm.037}$	$\underline{0.380}_{\pm.014}$	$0.336_{\pm.037}$	$0.398_{\pm.037}$
	LLM-KPs	$\underline{0.660}_{\pm.030}$	$0.362_{\pm.013}$	$\underline{0.641}_{\pm.031}$	$\underline{0.709}_{\pm.028}$
	LLM+DPP	$0.560_{\pm.035}$	$0.357_{\pm.011}$	$0.464_{\pm.030}$	$0.537_{\pm.032}$
	RICH	$\mathbf{0.720}_{\pm.037}$	$\mathbf{0.390}_{\pm.013}$	$\mathbf{0.686}_{\pm.031}$	$\mathbf{0.761}_{\pm.029}$
WCEP	Naive LLM	$0.680_{\pm.031}$	$0.357_{\pm.012}$	$0.290_{\pm.026}$	$0.335_{\pm.027}$
	LLM-KPs	$\underline{0.808}_{\pm.028}$	$\underline{0.381}_{\pm.007}$	$0.525_{\pm.027}$	$0.613_{\pm.025}$
	LLM+DPP	$0.768_{\pm.035}$	$0.374_{\pm.008}$	$\underline{0.539}_{\pm.032}$	$\underline{0.633}_{\pm.029}$
	RICH	$\mathbf{0.836}_{\pm.023}$	$\mathbf{0.387}_{\pm.010}$	$\mathbf{0.602}_{\pm.031}$	$\mathbf{0.684}_{\pm.028}$

on Answerability (κ=0.63) and 80% on Coverage (κ=0.60), with high inter-annotator agreement (Fleiss' κ=0.88). This supports the reliability of the automatic protocol (Appendix I).

Efficiency. Stage-by-stage timing on GPT-4o gives 41.0s per event end-to-end; RICH's own added compute (graph build plus recurrent GNN) is only $\approx$ 0.14s ($\approx 0.3\%$), the rest being the two LLM calls (extraction 27.8s, synthesis 13.1s) that every baseline also makes. This stage stays within 0.02–0.4s as the graph grows from 99 to 217 topics; the reported cost *includes* LLM topic extraction (Appendix G).

7 Conclusion

We presented RICH, a neuro-symbolic ranker for query-focused multi-document summarization that couples a vertex-reinforced Recurrent GNN over heterogeneous topic–document graphs with a self-supervised coverage–diversity objective. RICH leads strong baselines on Coverage and Semantic Spread across all four DiverseSumm setups with both backbones, and on Topic Cluster Coverage in every setup except the GPT-4o query-focused ones; the only places a baseline (LLM+DPP) edges ahead are GPT-4o Answerability and GPT-4o query-focused TCC, each by ≤ 0.01. The gains persist on an event-level held-out split with disjoint train/test events, significantly so in the query-focused settings, and transfer zero-shot to two unseen corpora. They stem from the bi-level structure, where an inner DivRank solver and an outer objective co-optimize the *same* learned graph topology, so diversity is baked into the representation rather than imposed post hoc (empirically it is driven mainly by the diversity term, Appendix J). Two limitations remain (*document resonance bias* and *analytical question mismatch*, Appendix O), to be addressed through multi-aspect intent expansion and adversarial rare-topic evaluation. Overall, principled content selection retains its value even as context windows expand.

Reproducibility Statement

We provide complete details to reproduce every result. The method, objectives, and coupled recurrent updates are specified in §4 with full training and inference pseudo-code in Appendix A. Every

hyperparameter, embedder, threshold, and optimizer setting used in all reported experiments is listed in Appendix D (Table 6), and the complete notation in Appendix E. Theoretical claims state their assumptions explicitly and are proved in full in Appendix B, with the convergence rate additionally measured on all evaluation graphs (Table 5). Dataset construction, the four evaluation setups, the leak-free training protocol, the event-level held-out split, and the metric definitions are described in §6.1; the human-validation protocol for the synthetic Q&A and the LLM judge is in Appendix I. All datasets used (DiverseSumm, Multi-News, WCEP) are public.

Ethics Statement

This work uses public news-summarization benchmarks and introduces no new human-subject data collection beyond a small expert annotation study for evaluation validation (Appendix I), conducted by consenting expert annotators on public news content. Structured multi-document summarization can help researchers, journalists, and knowledge workers surface diverse perspectives from large corpora. The same technology could be misused to produce digests that selectively omit inconvenient sources or amplify a narrative; RICH's explicit multi-source coverage objective and stochastic-dominance guarantee are architectural mitigations against single-source bias. We discuss broader impact and failure modes in Appendix O.

Use of Large Language Models

Consistent with the ICLR AI policy, we disclose that LLMs are integral *components* of the studied system: they perform atomic topic extraction and final summary synthesis (§4), generate the augmented Q&A and synthetic intents used for evaluation, and serve as the automatic judge (§6.1); all such uses are validated by the human study in Appendix I. Separately, generative-AI tools were used only to aid manuscript editing and to assist with routine code and experiment scaffolding; all research ideas, theoretical results, experimental design, and analyses are the authors' own, and the authors verified all AI-assisted text and code and take full responsibility for the content of this paper.

References

Jaime Carbonell and Jade Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 335–336, 1998.

Parthiv Chatterjee, Asish Joel Batha, Sourish Dasgupta, and Tanmoy Chakraborty. PerDucer: Keyphrase-driven personalization inducer for summarization from user histories. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 20651–20677. Association for Computational Linguistics, jul 2026.

Xueyu Chen, Kaitao Song, Zifan Song, Dongsheng Li, and Cairong Zhao. Improving long-context summarization with multi-granularity retrieval optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 30315–30323, 2026a. doi: 10.1609/aaai.v40i36.40283.

Yilei Chen, Souradip Chakraborty, Lorenz Wolf, Yannis Paschalidis, and Aldo Pacchiano. Post-training large language models for diverse high-quality responses. In *International Conference on Learning Representations (ICLR)*, 2026b.

Zhuo Chen, Oriol Mayné i Comas, Zhuotao Jin, Di Luo, and Marin Soljačić. L-CUBE: Isolating long-context capacity from knowledge with controllable mutual information scaling. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026c.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. From local to global: A graph RAG approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*, 2024.

Günes Erkan and Dragomir R Radev. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, 22:457–479, 2004.

Alexander R. Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1074–1084, 2019.

Demian Gholipour Ghalandari, Chris Hokamp, Nghia The Pham, John Glover, and Georgiana Ifrim. A large-scale multi-document summarization dataset from the Wikipedia current events portal. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1610–1625, 2020.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. News summarization and evaluation in the era of gpt-3. *arXiv preprint arXiv:2209.12356*, 2022.

Xiaoxin He, Yijun Tian, Yifei Sun, Nitesh V. Chawla, Thomas Laurent, Yann LeCun, Xavier Bresson, and Bryan Hooi. G-Retriever: Retrieval-augmented generation for textual graph understanding and question answering. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Cheng-Yu Hsieh, Yung-Sung Chuang, Chun-Liang Li, Zifeng Wang, Long T. Le, Abhishek Kumar, James Glass, Alexander Ratner, Chen-Yu Lee, Ranjay Krishna, and Tomas Pfister. Found in the middle: Calibrating positional attention bias improves long context utilization. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024.

Kung-Hsiang Huang, Philippe Laban, Alexander R. Fabbri, Prafulla Kumar Choubey, Shafiq Joty, Caiming Xiong, and Chien-Sheng Wu. Embrace divergence for richer insights: A multi-document summarization benchmark and a case study on summarizing diverse information from news articles. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 570–593, Mexico City, Mexico, jun 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.32.

Frederic Kirstein, Terry Ruas, Robert Kratel, and Bela Gipp. Tell me what I need to know: Exploring LLM-based (personalized) abstractive multi-source meeting summarization. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP): Industry Track*, 2024.

Alex Kulesza and Ben Taskar. Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning*, 5(2–3):73–286, 2012.

Philippe Laban, Tobias Schnabel, Paul N Bennett, and Marti A Hearst. Summac: Re-visiting nli-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177, 2022.

Yujia Li, Daniel Tarlow, Marc Brockschmidt, and Richard Zemel. Gated graph sequence neural networks. In *International Conference on Learning Representations (ICLR)*, 2016.

Dongqi Liu et al. What is that talk about? a video-to-text summarization dataset for scientific presentations. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics (TACL)*, 12:157–173, 2024.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2511–2522, 2023.

Himanshu Maheshwari, Sambaran Bandyopadhyay, Aparna Garimella, and Anandhavelu Natarajan. Presentations are not always linear! GNN meets LLM for text document-to-presentation transformation with attribution. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 15948–15962, 2024.

Qiaozhu Mei, Jian Guo, and Dragomir Radev. Divrank: the interplay of prestige and diversity in information networks. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1009–1018, 2010.

Rada Mihalcea and Paul Tarau. TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 404–411, 2004.

Vishakh Padmakumar, Zichao Wang, David Arbour, and Jennifer Healey. Principled content selection to generate diverse and personalized multi-document summaries. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, 1999.

Robin Pemantle. Vertex-reinforced random walk. *Probability Theory and Related Fields*, 92(1): 117–136, 1992.

Robin Pemantle. A survey of random processes with reinforcement. *Probability Surveys*, 4:1–79, 2007.

Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80, 2009.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations (ICLR)*, 2018.

Zhishang Xiang, Chuanjie Wu, Qinggang Zhang, Shengyuan Chen, Zijin Hong, Xiao Huang, and Jinsong Su. When to use graphs in RAG: A comprehensive analysis for graph retrieval-augmented generation. In *International Conference on Learning Representations (ICLR)*, 2026.

Yujia Xiao et al. PodAgent: A comprehensive framework for podcast generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, 2025.

Yue Yu, Wei Ping, Zihan Liu, Boxin Wang, Jiaxuan You, Chao Zhang, Mohammad Shoeybi, and Bryan Catanzaro. RankRAG: Unifying context ranking with retrieval-augmented generation in LLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2023.

APPENDIX

The appendix is organized as follows. Appendix A gives the training and inference algorithms. Appendix B contains the full theoretical development and all proofs (including the existence-and-stability Theorem 2 and the measured convergence rate), Appendices C–B.3 give the DPP-loss and hierarchical-prestige details. Appendix D lists every hyperparameter and Appendix E the notation. Appendices F–J contain the additional empirical results (out-of-distribution significance tests, latency and scalability, the full event-level held-out study, the human evaluation, and document-coverage-at-top-k); the main-text tables (Llama-3.1-8B and GPT-4o backbones, event-level held-out, and out-of-distribution transfer) are in §6.2. Appendix K reports the ablations, Appendices L–M describe two extensions and an adversarial benchmark, Appendix N presents case studies, and Appendix O discusses limitations and broader impact.

A ALGORITHMS

Algorithm 1 gives RICH's self-supervised bi-level training procedure and Algorithm 2 the end-to-end inference pipeline (stages S1–S4 and S6 of Figure 1, with S5 active only during training).

Algorithm 1 RICH Training (Self-Supervised Bi-Level Optimization)

Require: Samples $\{(\mathcal{D}_j, I_j)\}_{j=1}^{N}$; max epochs E; learning rate η; loss weights γ_1, γ_2; patience P
Ensure: Trained parameters Θ^*

Initialize Θ
Pre-processing (cached): for each $(\mathcal{D}_j, I_j)$: extract atomic topics $\mathcal{V}_j$ + parent map $a(\cdot)$ via LLM, embed $\{x_v\}$, compute document centroids $\{\mathbf{x}_{a_i}\}$, build hierarchical graphs G_t^j, G_a^j; encode intent x_{I_j} (or centroid if I_j=`None`) $\rightarrow$ cache
$\ell^* \leftarrow \infty$; $w \leftarrow 0$ ▷ best validation loss; stall counter
for $e = 1$ **to** E **do**
  for each cached $(G_t^j, G_a^j, \{x_v\}, \{\mathbf{x}_{a_i}\}, x_{I_j})$ **in shuffled order do**
    Compute priors (p^*, p_a^*) and GAT transitions (p_0, p_0^a) with Θ; run coupled recurrent updates for T steps (Eqs. 2–4) $\rightarrow (\pi_a^{(T)}, \mathbf{h}^{(T)})$
    $z_{\text{sum}} \leftarrow \sum_v h_v^{(T)} x_v$; $z_{\text{target}} \leftarrow \sum_v w_v x_v / \sum_v w_v$, $w_v = \max(0, \cos(x_v, x_{I_j}))$
    $\mathcal{L}_{\text{cov}} \leftarrow 1 - \cos(z_{\text{target}}, z_{\text{sum}})$; $L_\Theta \leftarrow \text{diag}\big(\sqrt{h^{(T)}}\big) K_\Theta \, \text{diag}\big(\sqrt{h^{(T)}}\big)$; $\mathcal{L}_{\text{DPP}} \leftarrow -\log\det(L_\Theta + \varepsilon I)$
    $\mathcal{L}_{\text{sp}} \leftarrow -\frac{1}{\log|\mathcal{V}|} \sum_v h_v^{(T)} \log(h_v^{(T)} + \epsilon)$; $\mathcal{L}_{\text{total}} \leftarrow \mathcal{L}_{\text{cov}} + \gamma_1 \mathcal{L}_{\text{DPP}} + \gamma_2 \mathcal{L}_{\text{sp}}$
    $\Theta \leftarrow \Theta - \eta \nabla_\Theta \mathcal{L}_{\text{total}}$ ▷ BPTT through coupled recurrent updates
  end for
  if validation loss $< \ell^*$ **then**
    $\ell^* \leftarrow$ val. loss; $\Theta^* \leftarrow \Theta$; $w \leftarrow 0$
  else
    $w \leftarrow w+1$; **if** $w \geq P$ **then break** ▷ early stopping
  end if
end for
return Θ^*

Algorithm 2 RICH Inference Pipeline

Require: Documents $\mathcal{D}=\{d_1, \ldots, d_K\}$; intent I (or `None`); parameters $\Theta=\{\mathbf{W}, \mathbf{a}, \mathbf{W}_a, \mathbf{a}_a, \text{MLP}, \text{MLP}_a\}$; $\lambda, \lambda_a, \alpha, T, \tau_{\text{extract}}, \tau_{\text{graph}}, \tau_a, k_{\text{nn}}$
Ensure: Summary s

S1. Extract atomic topics + parent map $a(\cdot)$ from each $d_i \in \mathcal{D}$ via LLM $\rightarrow (\mathcal{V}, a)$; embed $x_v \leftarrow \text{Encode}(v)$
if $I \neq$ `None` **then**
  $x_I \leftarrow \text{Encode}(I)$; $\mathcal{V} \leftarrow \{v : \cos(x_v, x_I) \geq \tau_{\text{extract}}\}$ ▷ intent filtering
else
  $x_I \leftarrow \frac{1}{|\mathcal{V}|} \sum_v x_v$ ▷ centroid as neutral intent
end if
Document centroids: $\mathbf{x}_{a_i} \leftarrow |\mathcal{V}_i|^{-1} \sum_{v:a(v)=a_i} x_v$ for each $a_i \in \mathcal{A}$
S2. Topic graph $G_t=(\mathcal{V}, E_t)$ with $E_t=\{(u,v) : \cos(x_u, x_v) \geq \tau_{\text{graph}} \vee v \in \text{kNN}(u, k_{\text{nn}})\}$; document graph $G_a=(\mathcal{A}, E_a)$ with $E_a=\{(a_i, a_j) : \cos(\mathbf{x}_{a_i}, \mathbf{x}_{a_j}) \geq \tau_a \vee k\text{-NN}\}$
S3a. Priors: $p^*(v) \leftarrow \text{softmax}(\text{MLP}([x_v \| x_I]))$, $p_a^*(a_i) \leftarrow \text{softmax}(\text{MLP}_a([\mathbf{x}_{a_i} \| x_I]))$ (PageRank fallback if I=`None`)
S3b. Transitions: GAT on $G_t \rightarrow p_0(u,v)$; GAT_a on $G_a \rightarrow p_0^a(a_i, a_j)$
S4. Initialize $\pi_a^{(0)} \leftarrow p_a^*$; $h_v^{(0)} \leftarrow \alpha\, p^*(v) + (1-\alpha)\, \pi_a^{(0)}(a(v))/|\mathcal{V}_{a(v)}|$
for $t = 0, \ldots, T-1$ **do**
  Doc update: $\pi_a^{(t+1)}(a_i) \leftarrow (1-\lambda_a) p_a^*(a_i) + \lambda_a \sum_{a_j \in \mathcal{N}(a_i)} \pi_a^{(t)}(a_j) \frac{p_0^a(a_j, a_i)\, \pi_a^{(t)}(a_i)}{\sum_{a_\ell} p_0^a(a_j, a_\ell)\, \pi_a^{(t)}(a_\ell)}$ ▷ Eq. 2
  Topic prior: $\tilde{p}^*(v) \leftarrow \alpha\, p^*(v) + (1-\alpha)\, \pi_a^{(t+1)}(a(v))/|\mathcal{V}_{a(v)}|$ ▷ Eq. 3
  Topic update: $h_v^{(t+1)} \leftarrow (1-\lambda)\, \tilde{p}^*(v) + \lambda \sum_{u \in \mathcal{N}(v)} \frac{p_0(u,v)\, h_v^{(t)}}{\sum_k p_0(u,k)\, h_k^{(t)}}\, h_u^{(t)}$ ▷ Eq. 4
end for
S6. Sort $\mathcal{V}$ in descending order of $h_v^{(T)}$ and emit ranked tuples $\mathcal{R} = \big\{(v, \text{evidence}(v), h_v^{(T)})\big\}_{v \in \mathcal{V}}$
return $s \leftarrow \text{LLM}(\mathcal{R})$

B Further Details on Theoretical Analysis

This appendix proves the three guarantees of §5: Theorem 1 (RICH is differentiable Pointwise DivRank, §B.1), Theorem 2 (existence and stability of the recurrence, §B.2) together with the

measured convergence rate, and Theorem 3 (per-source prestige floor, §B.4). The statements are deliberately elementary: each follows from the damped, mass-preserving structure of the DivRank update, and we mark exactly which assumptions each step uses.

B.1 Proof of Theorem 1 (RICH is Pointwise DivRank made differentiable)

The proof has two short parts: the update is term-for-term the Pointwise DivRank recurrence, and the resulting finite unrolling is differentiable, so gradients reach Θ by BPTT.

Part I: the recurrence is Pointwise DivRank. Mei et al. (2010) define Pointwise DivRank (their Eq. 10–11) by the reinforced walk

$$p_{T+1}(v) \;=\; (1-\lambda)\, p^*(v) \;+\; \lambda \sum_u p_0(u,v)\, \frac{p_T(v)}{D_T(u)}\, p_T(u), \qquad D_T(u) = \sum_v p_0(u,v)\, p_T(v), \tag{6}$$

where p^* is any prior distribution over vertices and p_0 is the organic (row-stochastic) transition kernel. Substituting $h^{(t)}$ for p_T, the learned GAT kernel p_0 for the organic kernel, and the hierarchical prior $\tilde{p}^*$ for p^*, Eq. 6 becomes, term by term, the RICH topic update (Eq. 4); the message denominator $\sum_{k\in\mathcal{N}(u)} p_0(u,k)h_k$ is exactly $D_T(u)$. So RICH is Pointwise DivRank with the transition kernel and the prior made learnable. The prior is admissible: DivRank allows any prior distribution, and $\tilde{p}^*$ is one, since dividing the parent prestige by $|\mathcal{V}_{a(v)}|$ gives

$$\sum_v \tilde{p}^*(v) = \alpha \sum_v p^*(v) + (1-\alpha) \sum_i \sum_{v\in\mathcal{V}_i} \frac{\pi_a(a_i)}{|\mathcal{V}_i|} = \alpha + (1-\alpha) \sum_i \pi_a(a_i) = 1. \tag{7}$$

Part II: differentiability and BPTT. Fix Θ. One step of RICH is a composition of the softmax priors, the GAT attention $p_0(u,v) = \mathrm{softmax}_v(\mathrm{LeakyReLU}(\cdots))$, and the reinforced update, which uses only sums, products, and divisions by the denominators $D_u(h) = \sum_k p_0(u,k)h_k$. By Theorem 2 the iterates keep every coordinate above the floor $\delta > 0$, so $D_u(h) \geq \delta \min_{u,k} p_0(u,k) > 0$ and no division is singular. Every operation is then C^∞ except the LeakyReLU, which is smooth off the measure-zero set where its argument vanishes. Hence the T-step composition $h^{(T)}(\Theta)$ is differentiable in Θ on a full-measure set, and backpropagation through the unrolled steps returns, by the chain rule,

$$\frac{d\, h^{(T)}}{d\Theta} \;=\; \sum_{t=1}^{T} \left(\prod_{i=t}^{T-1} \frac{\partial \Phi(h^{(i)}, \Theta)}{\partial h^{(i)}} \right) \frac{\partial \Phi(h^{(t-1)}, \Theta)}{\partial \Theta}, \tag{8}$$

so $\nabla_\Theta \mathcal{L}_{\text{total}}(h^{(T)})$ is well defined wherever $h^{(T)}$ is differentiable. RICH therefore trains end to end: the inner loop is the now-differentiable DivRank solver and the outer loop optimises Θ. ■

B.2 Proof of Theorem 2 (Existence and Stability)

Setup. RICH runs two prestige updates together. The document prestige π (a distribution over documents) follows a DivRank update on G_a and depends only on π; the topic prestige h (a distribution over topics) follows Eq. 4 with prior $\tilde{p}^*(\pi)$ from Eq. 3. Writing $z = (\pi, h)$, the joint update is $\Phi(z) = (\Psi(\pi), \Phi_h(h;\pi))$ with

$$\Psi(\pi) = (1-\lambda_a)\, p_a^* + \lambda_a\, W_a(\pi), \qquad \Phi_h(h;\pi) = (1-\lambda)\, \tilde{p}^*(\pi) + \lambda\, W(h), \tag{9}$$

where $W(h)_v \;=\; \sum_{u\in\mathcal{N}(v)} p_0(u,v)h_v\, h_u / \sum_{k\in\mathcal{N}(u)} p_0(u,k)h_k$ is the reinforced-walk step (and W_a likewise). The coupling is one-way: π evolves on its own and h depends on both. We use three assumptions, all holding *by construction* in RICH:

- **(A1) Connectivity and undirected graphs.** The k-NN step (k=5 topics, k=4 documents) is symmetrised and a self-loop is added to any isolated node, so both graphs are connected and undirected, the neighbourhoods $\mathcal{N}$ are symmetric, and p_0, p_0^a are row-stochastic.
- **(A2) Damping and strictly positive priors.** $\lambda, \lambda_a \in (0,1)$, and the priors p^*, p_a^* are softmax outputs, hence strictly positive.
- **(A3) Bounded fan-in.** Each document has at most N child topics.

Proof of (i): the simplex is invariant. The walk step conserves mass. Because $\mathcal{N}$ is symmetric (A1) and $\sum_{v\in\mathcal{N}(u)} p_0(u,v)h_v / \sum_{k\in\mathcal{N}(u)} p_0(u,k)h_k = 1$,

$$\sum_v W(h)_v = \sum_v \sum_{u\in\mathcal{N}(v)} \frac{p_0(u,v)h_v}{\sum_k p_0(u,k)h_k} h_u = \sum_u h_u \sum_{v\in\mathcal{N}(u)} \frac{p_0(u,v)h_v}{\sum_k p_0(u,k)h_k} = \sum_u h_u, \quad (10)$$

so $W(h)$ is a probability vector whenever h is. The prior $\tilde{p}^*$ is a probability vector (Eq. 7), so $\Phi_h(h;\pi) = (1-\lambda)\tilde{p}^* + \lambda W(h)$, a convex combination of two probability vectors, is again one; the same holds for Ψ using $\sum_i p_a^*(a_i) = 1$. Hence Φ maps the product of simplices into itself.

Proof of (ii): a positive floor. Dropping the non-negative walk term,

$$h_v^{(t+1)} \;\geq\; (1-\lambda)\,\tilde{p}^*(v) \;\geq\; (1-\lambda)\,\alpha\,p^*(v) \;\geq\; (1-\lambda)\,\alpha \min_v p^*(v) \;=:\; \delta \;>\; 0, \quad (11)$$

using $\alpha > 0$ and the strict positivity of the softmax prior (for $\alpha = 0$ the same bound follows from the document floor and A3). Likewise $\pi_{a_i}^{(t+1)} \geq (1-\lambda_a)\min_i p_a^*(a_i) =: \delta_a > 0$. Since these lower bounds do not depend on the current iterate, the floored simplices $\{h_v \geq \delta\}$ and $\{\pi_{a_i} \geq \delta_a\}$ are invariant.

Proof of (iii): existence and well-posed gradients. On the floored sets every denominator satisfies $\sum_k p_0(u,k)h_k \geq \delta \min_{u,k} p_0(u,k) > 0$, so Φ is continuous (indeed smooth off the $\mathrm{LeakyReLU}$ kinks). The floored product simplex is compact, convex, non-empty ($\delta\, n < 1$ since $\delta \leq 1/n$), and invariant, so Φ has a fixed point (π^*, h^*) by Brouwer's theorem. Positive denominators make the unrolled solver and its BPTT gradients (Eq. 8) well defined. ■

Convergence rate, measured. Theorem 2 gives existence and stability, not a rate: Brouwer yields a fixed point but not that the iterates reach it. We therefore measure the rate directly on all 245 evaluation graphs (Table 5): the per-step change $d_{\mathrm{TV}}(h^{(t+1)}, h^{(t)})$ shrinks by a roughly constant factor ≈ 0.73 per step, so the iteration contracts geometrically in practice and $T{=}10$ leaves the state within ≈ 0.008 (TV) of $h^{(T)}$. This is an empirical measurement of the rate, not a closed-form bound.

Table 5: Measured convergence of the coupled RecGNN, averaged over all 245 DiverseSumm evaluation graphs. The per-step change $d_{\mathrm{TV}}(h^{(t+1)}, h^{(t)})$ shrinks by a roughly constant factor each step; a geometric fit gives $\rho \approx 0.73$, well below 1. The last column reports distance to the final iterate $h^{(T)}$ (zero at $t{=}T$ by definition), not to the true fixed point. With $T{=}10$ the per-step change is already ≈ 0.008 (TV).

step t	mean $d_{\mathrm{TV}}(h^{(t+1)}, h^{(t)})$	mean $d_{\mathrm{TV}}(h^{(t)}, h^{(T)})$	ratio to previous
0	0.0441	0.1687	–
1	0.0306	0.1309	0.78
2	0.0230	0.1036	0.79
3	0.0184	0.0824	0.80
4	0.0152	0.0649	0.79
5	0.0130	0.0502	0.77
6	0.0113	0.0374	0.75
7	0.0099	0.0262	0.70
8	0.0087	0.0164	0.62
9	0.0077	0.0077	0.47
10	–	0.0000	–

The behaviour is the same for every event, with small spread across graphs, so on the real graphs the iteration contracts geometrically and $T{=}10$ is comfortably sufficient. We report this as a measured property, not a proved bound.

Practical implication for hyperparameters. The observed rate depends on λ (damping) and the MLP prior sharpness (which sets the floor δ). The floor can be raised by a small uniform mixture $\tilde{p}^*(v) \leftarrow (1-\varsigma)\,\tilde{p}^*(v) + \varsigma/n$ for $\varsigma \in [0.01, 0.05]$, which speeds up convergence with negligible effect on ranking quality.

B.3 Hierarchical Document-to-Topic Prestige and Product-Space Reduction

The flat RICH computes a single topic-level prestige $h^* \in \Delta^{|\mathcal{V}|-1}$, which encourages diversity but gives no structural guarantee that the head of the ranking spans multiple documents; if several high-centrality topics come from one dominant document, the rewriter over-allocates attention to it and undercovers the rest. The hierarchical extension couples a document-level graph G_a with G_t.

Document graph and prior. With $\mathbf{x}_{a_i} = |\mathcal{V}_i|^{-1} \sum_{v \in \mathcal{V}_i} x_v$ and $\mathcal{V}_i = \{v : a(v) = a_i\}$, edges E_a connect pairs with $\cos(\mathbf{x}_{a_i}, \mathbf{x}_{a_j}) \geq \tau_a$ (with k-NN fallback). An intent-conditioned document prior $p_a^*(a_i) = \mathrm{softmax}_i(\mathrm{MLP}_a([\mathbf{x}_{a_i} \| \mathbf{x}_I]))$ encodes document relevance, with MLP_a parameter-disjoint from the topic MLP.

Coupled updates. The document prestige follows the DivRank recurrence Eq. 2; the topic prior is augmented by the parent's running prestige via Eq. 3; and the topic prestige follows Eq. 4 using $\tilde{p}^*$. All other components (GAT, MLP, losses) are unchanged.

Product-space view. The joint prestige (π_a, h) can be viewed as a coupled walk over document–topic pairs $\{(a_i, v) : v \in \mathcal{V}_i\}$, whose kernel combines the two levels,

$$P\big((a_i, u) \to (a_j, v)\big) \;\propto\; p_0(u, v) \cdot \mathbf{1}[a_j = a(v)] \cdot p_0^a(a_i, a_j), \tag{12}$$

normalised to be row-stochastic, with the document prestige feeding the topic prior through Eq. 3 (a one-way coupling). Existence and stability of the joint fixed point are given by Theorem 2; we do not claim it equals a symmetric product-graph stationary distribution, since the coupling runs only from documents to topics.

B.4 Proof of Theorem 3 (Per-source prestige floor)

Restated. Under the hierarchical prior (Eq. 3) with $\alpha < 1$, every topic v satisfies, at every step t,

$$h_v^{(t+1)} \;\geq\; (1-\lambda)\,(1-\alpha)\,\frac{\pi_a^{(t)}(a(v))}{|\mathcal{V}_{a(v)}|} \;>\; 0.$$

Proof. The topic update is $h_v^{(t+1)} = (1-\lambda)\,\tilde{p}^*(v) + \lambda \sum_{u \in \mathcal{N}(v)} m_{u \to v}^{(t)}$ with every message $m_{u \to v}^{(t)} \geq 0$, so $h_v^{(t+1)} \geq (1-\lambda)\,\tilde{p}^*(v)$. By definition $\tilde{p}^*(v) = \alpha\, p^*(v) + (1-\alpha)\,\pi_a^{(t)}(a(v))/|\mathcal{V}_{a(v)}| \geq (1-\alpha)\,\pi_a^{(t)}(a(v))/|\mathcal{V}_{a(v)}|$, since $\alpha\, p^*(v) \geq 0$. Combining gives the bound, which is strictly positive because $\pi_a^{(t)}(a(v)) \geq \delta_a > 0$ (Theorem 2). The damping term $(1-\lambda)\tilde{p}^*$ is re-added at every step, so the floor holds for all t, including the final T. With $\alpha = 1$ the injected term vanishes and the floor is 0. ■

Scope and empirical attribution. This is a per-source floor, not a coverage-dominance claim: it guarantees each source a minimum presence but does not order the two rankings' coverage. Empirically, the broad multi-document coverage RICH attains comes mainly from the diversity (DPP) term, not the α-blend alone (Appendix J); the failure cases of Appendix O are where one dense document pair dominates despite the floor, so they lie outside this guarantee rather than contradict it. We state the theorem and the discussion consistently with this measurement.

C DPP Log-Likelihood Diversity Loss

A pairwise diversity penalty $\mathcal{L}_{\text{div}} = \sum_{u,v} h_u h_v \max(0, \cos(x_u, x_v))$ is a loose surrogate: it corresponds to no probabilistic diversity criterion and is only a first-order approximation of the volume spanned by the selected set. We replace it with the DPP log-likelihood loss, tied to the DPP literature (Kulesza & Taskar, 2012).

Kernel and L-ensemble. Let $h = h^{(T)}$ be the converged prestige. The GAT-learned similarity kernel is the symmetrised transition matrix $K_\Theta[u, v] = \frac{1}{2}(p_0(u, v) + p_0(v, u)) \in [0, 1]$. Following

the L-ensemble construction, $L_\Theta = \mathrm{diag}(\sqrt{h})\, K_\Theta\, \mathrm{diag}(\sqrt{h})$, so $L_\Theta[u,v] = \sqrt{h_u}\, K_\Theta[u,v]\, \sqrt{h_v}$: higher-prestige topics contribute larger diagonal entries (quality) and off-diagonals encode redundancy (repulsion). $\det(L_S)$ is large iff S is high-prestige and mutually dissimilar.

Loss and relation to $\mathcal{L}_{\text{div}}$. We set $\mathcal{L}_{\text{DPP}} = -\log\det(L_\Theta + \varepsilon I)$ (Eq. 5), $\varepsilon = 10^{-5}$ for positive-definiteness. The Maclaurin expansion $-\log\det(I+A) = -\operatorname{tr}(A) + \frac{1}{2}\operatorname{tr}(A^2) - \cdots$ gives, for small off-diagonals,

$$\mathcal{L}_{\text{DPP}} \approx -\sum_v \log h_v \;+\; \sum_{u \neq v} h_u h_v K_\Theta[u,v] \;+\; O(\|L_\Theta\|_F^2). \tag{13}$$

The second term is exactly $\mathcal{L}_{\text{div}}$; $\mathcal{L}_{\text{DPP}}$ additionally retains the log-diagonal coverage term $-\sum_v \log h_v$ that pushes h away from degenerate point masses, so it is the full principled objective.

Computation. We compute $-\log\det$ stably via Cholesky: for $M = L_\Theta + \varepsilon I = \mathbf{C}\mathbf{C}^\top$, $-\log\det M = -2\sum_i \log \mathbf{C}_{ii}$. With $|\mathcal{V}| \leq 250$, the Cholesky costs $O(|\mathcal{V}|^3) \approx 1.6 \times 10^7$ ops (< 1 ms) and is natively differentiable in PyTorch.

D Detailed Hyperparameter Setup for Reproducibility

Table 6 lists every hyperparameter used in the reported experiments.

Embedding and topic extraction. Topics are extracted per document with an instruction-following LLM and embedded with `intfloat/e5-large-v2` ($d = 1024$, Apache-2.0). In query-focused mode (S3–S4) a cosine filter $\tau_{\text{extract}} = 0.75$ discards topics far from the intent; generic mode uses all extracted topics.

Graph construction. Topic graph G_t: an edge is added when $\cos(x_u, x_v) \geq \tau_{\text{graph}}$, applied *adaptively* (per-event 25th percentile of the pairwise cosine distribution, hard minimum 0.65); a k-NN fallback ($k = 5$) ensures connectivity and isolated nodes receive self-loops. Document graph G_a: mean-pooled topic embeddings, edges require $\cos(\mathbf{x}_{a_i}, \mathbf{x}_{a_j}) \geq \tau_{\text{doc}} = 0.55$, $k_{\text{doc}} = 4$ fallback (the lower threshold reflects the broader cosine distribution of document-level embeddings).

Architecture. Topic GAT: single-head, projection width 128, LeakyReLU slope 0.2, row-stochastic output. Document GAT: same, width 64 (≤ 10 nodes per event). Topic MLP: two layers, input $2d = 2048$, hidden 256, softmax over $|\mathcal{V}|$. Document MLP: separate two layers, hidden 256, softmax over $K \leq 10$. Total trainable parameters $\approx 2.8 \times 10^5$, at least three orders of magnitude below the embedding model and the LLM rewriter.

Recurrent updates and optimization. $T = 10$ unrolled steps; $\lambda = \lambda_a = 0.85$; $\alpha = 0.6$. The loss is $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cov}} + \gamma_1 \mathcal{L}_{\text{DPP}} + \gamma_2 \mathcal{L}_{\text{sparsity}}$ with $\gamma_1 = 0.5$, $\gamma_2 = 0.1$; DPP ridge $\varepsilon_{\text{DPP}} = 10^{-5}$, numerical floor $\varepsilon = 10^{-8}$. Training uses AdamW ($\eta = 10^{-3}$, weight decay 10^{-5}) for up to 15 epochs with early stopping (patience 3), 10% of the training pool held out for validation. To avoid label leakage, training uses only generic (centroid) intents and synthetic query-focused intents derived from article content, never the evaluation Q&A pairs; gradients propagate through the T=10 unrolled steps via BPTT.

Compute. All RICH training and inference ran on a CPU-only machine (Intel Xeon Platinum 8375C @ 2.90 GHz, 96 vCPUs, ≈186 GB RAM; no GPU). Training converges in ≤15 epochs in ≈2 hours per run. Inference cost is dominated by the single LLM synthesis call; baselines required comparable LLM budgets.

E Notation

F Out-of-Distribution Significance Tests

This appendix reports the significance tests for the zero-shot out-of-distribution transfer to Multi-News (Fabbri et al., 2019) and WCEP (Gholipour Ghalandari et al., 2020) presented in Table 4 of the

Table 6: Complete hyperparameter listing for RICH as used in all reported experiments.

Group	Parameter	Symbol	Value
Embedding	Model	–	`intfloat/e5-large-v2`
	Output dimension	d	1024
	Extraction filter	τ_{extract}	0.75 (query-focused only)
Topic graph G_t	Similarity threshold	τ_{graph}	0.65 (adaptive 25th percentile)
	k-NN fallback	k	5
	Self-loops (isolates)	–	enabled
	Threshold mode	–	adaptive
Doc graph G_a	Similarity threshold	τ_{doc}	0.55
	k-NN fallback	k_{doc}	4
Architecture	Topic GAT hidden dim	–	128
	Document GAT hidden dim	–	64
	MLP hidden dim (both)	–	256
	LeakyReLU slope	–	0.2
RecGNN	Recurrent steps	T	10
	Topic damping	λ	0.85
	Document damping	λ_a	0.85
	Prior blend	α	0.6
Loss	DPP weight	γ_1	0.5
	Sparsity weight	γ_2	0.1
	DPP ridge	ε_{DPP}	10^{-5}
	Numerical floor	ε	10^{-8}
	DPP kernel	–	cosine (normalised by $\lvert\mathcal{V}\rvert$)
Optimisation	Optimizer	–	AdamW
	Learning rate	η	10^{-3}
	Weight decay	–	10^{-5}
	Max epochs	E	15
	Early-stop patience	P	3 epochs
	Validation fraction	–	10%

main text (25 events per dataset, Setup-2 protocol: generic summaries, GPT-5.1-generated augmented Q&A, Llama-3.1-8B backbone, GPT-5.1 judge; RICH is evaluated with the DiverseSumm-trained checkpoint and no fine-tuning). The consistent TCC/SS improvements in Table 4 confirm the diversity gains transfer, validating the generality of the learned graph topology rather than memorization of DiverseSumm-specific patterns.

Significance. We report a paired bootstrap (10,000 resamples) and paired t-tests. On Multi-News, RICH beats Naive LLM and LLM+DPP at $p < 0.001$ on both Coverage and Answerability, and LLM-KPs at $p = 0.020$ (Coverage) / $p = 0.005$ (Answerability). On WCEP, RICH beats Naive LLM and LLM-KPs at $p < 0.001$ on both metrics, and LLM+DPP at $p = 0.001$ (Coverage) / $p = 0.010$ (Answerability). On the 30-event ablation (Appendix K) most pairwise differences are not significant at this sample size, so we report the ablation as a directional trend.

G Latency and Scalability

We added stage timers and measured time per event, separating the two LLM stages (S1 extraction, S6 synthesis) from RICH's new local compute (S2 graph build, S3–S4 recurrent GNN). Live end-to-end timing on the GPT-4o backbone is in Table 8.

Scalability. The graph-plus-GNN stage stays in the 0.02–0.4s range as the topic graph grows from 99 to 217 topics, so it is not a bottleneck at these sizes; end-to-end time is set by the two LLM calls. On a warm cache (topic embeddings via `e5-large-v2`, no LLM), the local novel compute is ≈ 0.09s per event over 28 events ($|\mathcal{V}| = 99$–201). Full-dataset log mining shows RICH's total

Table 7: Summary of notation used throughout the paper.

Symbol	Description	Symbol	Description
Problem setup			
$\mathcal{D} = \{d_1, \ldots, d_K\}$	Collection of K source documents	I	User intent (personalized information need)
K	Number of source documents	s	Generated summary
$\mathcal{V} = \{v_1, \ldots, v_n\}$	Set of n atomic topics extracted from $\mathcal{D}$	$\mathcal{A} = \{a_1, \ldots, a_K\}$	Set of source document indices
$a : \mathcal{V} \to \mathcal{A}$	Parent map: source document of each topic	$\mathcal{V}_i = \{v : a(v) = a_i\}$	Topics belonging to document a_i
$\pi : \mathcal{V} \to [0, 1]$	Final ranking function over topics	evidence(v)	Supporting evidence span(s) for topic v
Embeddings			
d	Embedding dimension (1 024)	$x_v \in \mathbb{R}^d$	Dense embedding of topic v
$x_I \in \mathbb{R}^d$	Dense embedding of user intent I	$\mathbf{x}_{a_i} \in \mathbb{R}^d$	Document centroid: $\lvert\mathcal{V}_i\rvert^{-1} \sum_{v \in \mathcal{V}_i} x_v$
Graphs and thresholds			
$G_t = (\mathcal{V}, E_t)$	Topic-level semantic graph	$G_a = (\mathcal{A}, E_a)$	Document-level semantic graph
$\mathcal{N}(v)$	Neighbours of topic v in G_t	$\mathcal{N}(a_i)$	Neighbours of document a_i in G_a
τ_{graph}	Topic graph similarity threshold (0.65)	τ_{doc}	Document graph similarity threshold (0.55)
τ_{extract}	Intent-extraction cosine filter (0.75)	k, k_{doc}	k-NN fallback degrees (5, 4)
Neural parameters (Θ)			
$\mathbf{W}, \mathbf{a}$	Topic GAT weight matrix and attention vector	$\mathbf{W}_a, \mathbf{a}_a$	Document GAT weight matrix and attention vector
MLP	Topic-level intent-prior MLP	MLP_a	Document-level intent-prior MLP
Θ	All trainable parameters $\{\mathbf{W}, \mathbf{a}, \mathbf{W}_a, \mathbf{a}_a, \text{MLP}, \text{MLP}_a\}$	N	Bounded fan-in (max child topics per doc)
Priors and GAT transition kernels			
$p^*(v)$	Topic-level intent prior	$p^*_a(a_i)$	Document-level intent prior
e_{uv}	Unnormalised GAT attention score	$p_0(u, v)$	Topic-level row-stochastic transition kernel
$p^a_0(a_i, a_j)$	Document-level row-stochastic transition kernel	$\tilde{p}^*(v)$	Document-augmented topic prior (Eq. 3)
RecGNN dynamics			
$h^{(t)}_v$	Topic prestige (hidden state) at step t	$\pi^{(t)}_a(a_i)$	Document prestige at step t
$m^{(t)}_{u \to v}$	Vertex-reinforced message from u to v	$h^* = h^{(T)}$	Converged topic prestige distribution
$\lambda \in (0, 1)$	Topic-level damping (0.85)	$\lambda_a \in (0, 1)$	Document-level damping (0.85)
$\alpha \in [0, 1]$	Prior blend coefficient (0.6)	T	Number of RecGNN unrolling steps (10)
Φ, Φ_h, Ψ	Fixed-point update maps (joint / topic / doc)	δ, δ_a	Derived prestige floors (topic / doc)
ρ	Measured convergence rate (≈ 0.73, Table 5)		
Training objectives			
z_{sum}	Predicted summary embedding	z_{target}	Intent-targeted embedding
w_v	Intent-alignment weight $\max(0, \cos(x_v, x_I))$	$K_\Theta[u, v]$	Symmetrised GAT similarity kernel
L_Θ	DPP quality-diversity L-ensemble	$\mathcal{L}_{\text{cov}}$	Intent coverage loss
$\mathcal{L}_{\text{DPP}}$	DPP log-likelihood diversity loss	$\mathcal{L}_{\text{sparsity}}$	Sparsity regularizer
$\mathcal{L}_{\text{total}}$	Total training loss	γ_1, γ_2	Loss weights (0.5, 0.1)
ε	DPP ridge (10^{-5})	$\Delta^{\lvert\mathcal{V}\rvert - 1}$	Probability simplex over $\mathcal{V}$
Evaluation			
Q	Set of evaluation questions for an event	q, a	Evaluation question and reference answer
$C(k)$	Document-coverage fraction at ranking depth k		

generation time ($\approx$ 7534s over 245 events) is comparable to the strongest LLM baseline ($\approx$ 7226s for LLM+DPP), confirming RICH adds no dominant overhead.

Clarification of "negligible overhead". The end-to-end cost includes both LLM calls (topic extraction and synthesis); "negligible overhead" refers specifically to RICH's graph and GNN stage (0.14s). The LLM key-point extraction is per document and can be cached (the implementation keeps a disk cache keyed by document text and intent; the numbers above use a cold cache, i.e. the worst case), so extraction can be computed once per collection and reused across queries. Extraction is also not specific to RICH: LLM-KPs and LLM+DPP extract key points with an LLM too, so it is a cost shared by all three methods.

Table 8: Stage-by-stage latency per event (live, GPT-4o backbone). RICH's own new compute (graph build + recurrent GNN) is ≈ 0.14s, about 0.3% of the 41.0s end-to-end cost. The remaining 99.7% is the two LLM calls, which every LLM baseline also makes. The reported end-to-end cost *includes* LLM topic extraction (S1).

Stage	seconds per event
S1 Topic extraction (LLM)	27.78 ± 6.98
S2 Graph build (CPU)	0.005 ± 0.001
S3–S4 Recurrent GNN (CPU)	0.14 ± 0.11
S6 Synthesis (LLM)	13.08 ± 1.57
End-to-end total	$\mathbf{41.0 \pm 7.0}$

H Full Event-Level Held-Out Study

Table 3 (main text) reports the event-level held-out results with the Llama-3.1-8B backbone across all four setups. Here we give the full protocol and the complementary GPT-4o generic results.

Protocol. We create a single fixed split of the 245 DiverseSumm events into 196 training and 49 held-out events (sorted event ids, every 5th held out), with no overlap ($196 + 49 = 245$; train $\cap$ held-out $= \emptyset$). We retrain the RICH ranker on the 196 training events and evaluate on the 49 held-out events. The baselines do not train on events, so we take their scores on the same 49 events. The judge (GPT-5.1) and the metric protocol match the main tables. This is a strictly more demanding test than the leak-free protocol of §6.1 (which holds out only the evaluation Q&A): here the ranker never sees the held-out events at all.

Reproduction reference (Llama-3.1-8B). The held-out-trained model reproduces the reported model (trained on all 245 events) evaluated on the same 49 events. The reported model's Coverage / Answerability on these 49 events are S1 0.589 / 0.707, S2 0.457 / 0.571, S3 0.443 / 0.540, S4 0.445 / 0.556; the held-out-trained values in Table 3 match these to within ≈ 0.00–0.06 Coverage per setup (e.g. S2 0.456 vs. 0.457; S3 0.444 vs. 0.443; S4 0.460 vs. 0.445). The ranker therefore transfers to events it never saw, rather than overfitting the training fold.

GPT-4o generic held-out. Table 9 reports the two generic setups with the GPT-4o backbone. The held-out-trained RICH reproduces the reported model on the same 49 events, is well above Naive LLM and LLM-KPs, and is close to LLM+DPP in this generic regime. The stronger separation appears with the smaller backbone and in the query-focused setups (Table 3), where structured pre-selection matters most.

Table 9: Event-level held-out results, **GPT-4o** backbone, generic setups (49 held-out events, GPT-5.1 judge). The held-out-trained RICH reproduces the reported (245-event) model on the same events.

	S1: generic, original		**S2: generic, augmented**	
Method	Coverage	Answerability	Coverage	Answerability
Naive LLM	0.597	0.811	0.575	0.840
LLM-KPs	0.680	0.795	0.608	0.847
LLM+DPP	0.794	0.912	0.759	0.928
RICH (retrained on 196)	0.789	0.896	0.751	0.903
RICH *reported (245), same 49*	*0.789*	*0.883*	*0.753*	*0.927*

Backbone-serving note. For the Llama-3.1-8B held-out run, the backbone (topic extraction and generation) was served locally with vLLM (`meta-llama/Llama-3.1-8B-Instruct`), while the GPT-5.1 judge and all reused baselines stayed on the same endpoint as the main tables. The local backbone was calibrated to reproduce the reported run's metric-relevant quantities (topics per event and summary length), and the locally served RICH is, if anything, marginally below its counterpart, so the reported RICH-over-LLM+DPP margins are conservative. Two events whose RICH generation prompts exceeded the context window (extreme topic counts) yielded empty summaries and were

excluded from *every* method for the affected setup (paired, n=48); since these were RICH-side failures, excluding them slightly raises RICH, so the direction of the result is unaffected.

I HUMAN VALIDATION OF THE EVALUATION PROTOCOL

Because the main evaluation uses GPT-5.1-generated Q&A, synthetic intents, and a GPT-5.1 judge, we ran a human study to validate these artefacts. Recent work shows strong LLM judges align well with human judgments (Zheng et al., 2023; Liu et al., 2023); GPT-5.1 is a more recent and more capable model of this kind, and the study below confirms the alignment on our task.

Annotators and sample. Three expert annotators (fluent in English, expert in AI) worked independently. From a fixed seed we drew a stratified sample of 95 items: 45 (event, question) pairs for question validity (Task A) and 50 (question, summary) pairs for judge agreement (Task B). Each annotator labeled 45 items; Task A had 9 shared items and Task B had 11 shared items, used to measure inter-annotator agreement.

Task A: question validity. Each item showed a question, its reference answer, and the full source article. Annotators marked whether the question is (a) answerable from the source and (b) on-topic and non-trivial; a question was accepted only if both. **43 of 45 questions (95.6%) were accepted**, so almost all synthetic questions are valid.

Task B: judge agreement. Each item showed a question and a machine-generated summary, *blind* to the GPT-5.1 verdict (kept in a separate key). Annotators marked Answerability and Coverage; we compared human labels with the hidden verdict on all 50 items (Table 10). Agreement is $\approx 80\%$ on both metrics, well beyond chance, and inter-annotator agreement is high (Fleiss' $\kappa = 0.88$), so the human labels are reliable and the judge tracks them.

Table 10: Human vs. GPT-5.1 judge agreement on 50 blind (question, summary) items.

Metric	Human vs. judge agreement	Cohen's κ
Answerability	82%	0.63
Coverage	80%	0.60

J DOCUMENT COVERAGE AT TOP-k

We compute $C(k)$, the fraction of an event's source documents appearing in the top-k ranked topics, directly from saved ranked lists with no new inference. Theorem 3 guarantees each source a positive score floor but does not order the top-k coverage of the two rankings, so we measure the effect of the α-blend on $C(k)$ directly here.

The α-blend alone. Comparing the full model ($\alpha = 0.6$) against the no-hierarchy ablation ($\alpha = 1$) on the 30 ablation events gives values that are close at all k and not significantly different (e.g. k=10: 0.573 vs. 0.558, $p = 0.67$; k=20: 0.776 vs. 0.799, $p = 0.39$). On this small set (n=30) the α-blend alone shows no significant per-sample coverage gap, consistent with Theorem 3 being a per-source floor rather than a coverage-dominance claim.

The diversity term. The diversity (DPP) term gives a clear coverage gain: the full model covers more source documents than the no-DPP ablation by $+0.09$ at k=10 ($p = 0.011$), $+0.20$ at k=15, and $+0.26$ at k=20 (both $p < 0.001$). The broad multi-document coverage RICH shows in practice therefore comes mainly from the diversity term, consistent with the document-resonance behaviour discussed in Appendix O.

K Model Ablation Studies

To isolate each architectural component we evaluate three ablations on a 30-event stratified subset (stride-8) of DiverseSumm with the Llama-3.1-8B backbone and GPT-5.1 judge: **A1** disables the hierarchical document prior (α=1); **A2** removes the DPP loss (γ_1=0); **A3** collapses the RecGNN to a single step (T=1). Table 11 reports all four setups.

Table 11: Model ablation results on a 30-event subset (Llama-3.1-8B backbone, GPT-5.1 judge). **Bold** = best per metric per setup; underline = second best. A1 removes the hierarchy (α=1); A2 removes the DPP loss (γ_1=0); A3 collapses the RecGNN to T=1.

Setup	Variant	TCC	SS	Coverage	Answerability
S1: Generic, Original Q&A	A1: w/o Hierarchy	**0.830**	0.390	0.492	0.667
	A2: w/o $\mathcal{L}_{\text{DPP}}$	0.817	**0.395**	0.619	0.697
	A3: Static RecGNN	0.803	0.382	0.586	0.681
	RICH (full)	0.817	0.391	**0.633**	**0.728**
S2: Generic, Augmented Q&A	A1: w/o Hierarchy	**0.830**	0.390	0.508	0.632
	A2: w/o $\mathcal{L}_{\text{DPP}}$	0.817	**0.395**	0.521	0.605
	A3: Static RecGNN	0.803	0.382	0.536	0.632
	RICH (full)	0.817	0.391	**0.552**	**0.659**
S3: Query-Focused, All Aug. Q&A	A1: w/o Hierarchy	0.823	0.384	**0.527**	**0.644**
	A2: w/o $\mathcal{L}_{\text{DPP}}$	**0.827**	0.385	0.493	0.608
	A3: Static RecGNN	0.807	**0.389**	0.486	0.598
	RICH (full)	0.807	0.379	0.525	0.633
S4: Query-Focused, Relevant Q&A	A1: w/o Hierarchy	0.823	0.384	**0.563**	**0.671**
	A2: w/o $\mathcal{L}_{\text{DPP}}$	**0.827**	0.385	0.504	0.635
	A3: Static RecGNN	0.807	**0.389**	0.518	0.625
	RICH (full)	0.807	0.379	0.544	0.640

Findings. A2 consistently reduces Coverage across setups; A3 lowers Coverage and Answerability in every setup and reduces TCC and SS in the generic setups; A1 causes the largest drop in generic Coverage (S1: -0.14), confirming each component contributes. In the query-focused setups A1 marginally leads the full model on this small sample, consistent with the document-resonance bias (Appendix O): with a strong query intent, the hierarchical prior can over-weight a single high-cosine document. All three variants are trained identically to the full model (AdamW, η=10^{-3}, up to 10 epochs, patience 3) on the same stride-8 subset.

Caveat on sample size. The 30-event subset is intentionally small to keep ablation compute tractable on CPU. With n=30 the per-cell standard error is $\approx \sigma/\sqrt{30}$, i.e. $\pm$0.03–0.04 for typical score standard deviations of 0.15–0.20; differences of this magnitude (particularly the A1 advantage in S3–S4) should not be over-interpreted. A paired bootstrap on this set finds only A1-vs-full on S1 Coverage reaching $p < 0.05$; we therefore treat the ablation as directional, with the primary component evidence coming from the full-scale comparisons in §6.2 and the top-k coverage analysis of Appendix J.

L Multi-Aspect Intent Expansion at Inference

A user's query I often encodes multiple aspects (e.g. "causes, timeline, and geopolitical consequences of the conflict" spans three sub-intents). The base RICH processes a single intent embedding, which may under-weight aspects that are lexically minor but thematically important. Multi-aspect expansion decomposes I into K_I sub-intents, runs the RecGNN per aspect, and aggregates.

Procedure. *(1) Decomposition:* a single LLM call yields $\{I_1, \ldots, I_{K_I}\}$; K_I=1 with I_1=I recovers the base method. *(2) Encoding:* $x_{I_j} = \text{Embed}(I_j)$. *(3) Per-aspect passes:* compute $p_j^*(v) = \text{softmax}_v(\text{MLP}([x_v \| x_{I_j}]))$ and run the T-step RecGNN to get $h^{(j)}$. *(4) Aggregation*, by either max-pooling (union semantics), $h_{\text{agg}}(v) = \max_j h^{(j)}(v) / \sum_u \max_j h^{(j)}(u)$, or a weighted average $h_{\text{agg}}(v) = \sum_j w_j h^{(j)}(v)$ with $\sum_j w_j = 1$ (uniform or $w_j \propto \cos(x_{I_j}, x_I)$). Topics are then sorted by h_{agg} and handed to the rewriter.

Cost. Decomposition adds one LLM call; each RecGNN pass costs $O(T|E_t|)$ (< 5 ms on CPU for typical events). With K_I=3, overhead is one extra LLM call plus $\sim$15 ms compute, dominated by the LLM call.

Training-inference mismatch. Inference-only expansion introduces distribution shift (the MLP was trained on full-sentence intents but receives shorter sub-intents), a loss mismatch (coverage/DPP losses were trained against single-intent $h^{(T)}$, and h_{agg} under max-pooling can be flatter), and a scaling issue (max-pooling breaks the simplex constraint, requiring renormalization). We recommend *mixed training with stochastic intent decomposition*: with probability $p_{\text{expand}} \in [0.3, 0.5]$, sample $K_I \sim \text{Uniform}\{1, \ldots, K_{\max}\}$, decompose, and evaluate the loss on h_{agg}; the rest of the time use K_I=1. This reduces the mismatches while retaining the single-intent baseline as a degenerate case.

M DiverseSumm-LongTail: An Adversarial-Intent Benchmark

The main setups use generic or event-topic intents, which tend to be answerable from many articles, making it hard to isolate genuine personalization from article-order position bias. We outline **DiverseSumm-LongTail**, a slice that stress-tests whether a method surfaces *rare, minority-source facts* that directly answer a personalized query.

Long-tail fact identification. For each event (K=10 articles), a long-tail fact appears in at most $\lfloor K/5 \rfloor = 2$ articles. The pipeline: (1) extract atomic facts F_i per article with the RICH extractor; (2) cluster facts across articles by paraphrase equivalence ($\cos \geq 0.85$); (3) retain clusters with article support ≤ 2, whose representative facts form the long-tail set $\mathcal{F}^{\text{LT}}$.

Targeted intent synthesis and validation. For each $f \in \mathcal{F}^{\text{LT}}$, an LLM writes a one-sentence query whose direct answer is f (without paraphrasing f), yielding an intent I_f answerable only from the 1–2 articles containing f. We keep I_f if $\cos(x_{I_f}, x_f) \in [0.50, 0.85]$ (loose enough to be non-trivial, tight enough to be relevant) and if an LLM judge finds it answerable from the fact-bearing article but not from a random non-fact-bearing article.

Scale and protocol. We expect $\approx$ 3–5 valid long-tail intents per event across 245 events ($\approx$ 735–1,225 intent–event pairs), denoted **Setup 5 (Long-Tail QF)**. Evaluation uses the same four-metric protocol as S3–S4; the critical metric is Answerability, since I_f is answerable only from minority-source articles, so a high score proves the method surfaced minority-source content against majority-source position bias.

Expected behaviour. Article-order methods should achieve only moderate Answerability (minority-source topics land at uniformly random prompt positions, disadvantaged by LLM position bias). Intent-agnostic DPP selection should drop further (its kernel spreads over all topics without knowing I_f targets a rare fact). RICH should achieve substantially higher Answerability, because the intent prior up-weights the rare-fact topic and the RecGNN amplifies it to the head of the ranking, where LLM attention is strongest, providing cleaner evidence for the personalization claim than the generic setups can.

N Case Studies

We present two success cases with strong multi-source coverage and two failure cases revealing recurring limitations, drawn from the generic-summary evaluation (Setup 2, Llama-3.1-8B) on DiverseSumm. For each event we report Coverage for RICH and LLM+DPP, the converged topic-score distribution, and rank shifts of the top topics.

N.1 Success Cases

Case S1: Neutrino Detection at the LHC. Ten articles on the first detection of collider-produced neutrinos by FASER at CERN ($|V|$=127, $|E|$=12,004). RICH achieves Coverage 0.909 vs. LLM+DPP 0.455 (Naive LLM 0.318). The initial PageRank prior is nearly uniform ($h^{(0)}_{\max}$=0.0103,

std 0.0018); after T=10 steps it is highly peaked ($h^*_{\max}$=0.386, std 0.042). The two highest-scoring final topics rose from initial ranks #108 and #76 (+107, +74): a cosmological-implication claim (doc 5) and a technical-context claim (doc 2), semantically distant from each other but each densely corroborated across the 10 documents. The vertex-reinforced walk amplifies them jointly without collapsing to a single document; the summary covers the discovery from five distinct angles. LLM+DPP's monolithic kernel misses the cross-document corroboration and reaches only 45.5%.

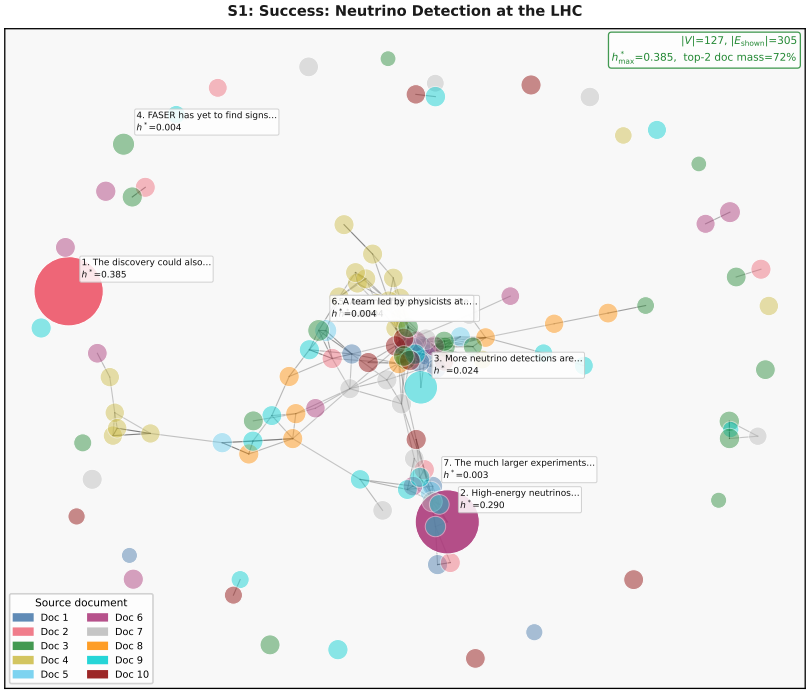

Figure 2: **Case S1: Neutrino Detection at the LHC.** Topic graph after RecGNN convergence ($|V|$=127, edges with similarity $\geq$0.88 shown). Node size/colour intensity encode h^*; colour encodes source document. The two highest-scoring topics rose by +107 and +74 positions.

Case S2: Canadian Inflation and Bank of Canada Rate Decision. Ten articles on Canada's February 2023 CPI release and the subsequent rate-hold ($|V|$=131, 10 docs). RICH achieves **perfect** Coverage 1.000 vs. LLM+DPP 0.000. The question probed "signs of stickiness in consumer prices," requiring synthesis across three articles (grocery prices above 10%, core inflation near 5%, and British Columbia inflation not cooling). These articles, each citing the same Statistics Canada release, form a *triangle of mutual reinforcement* the walk exploits; LLM+DPP, selecting for geometric spread, misses the corroboration and scores 0.

N.2 Failure Analysis

Case F1: Starbucks Labour Dispute (Document Resonance Bias). Ten articles on Starbucks union activity ($|V|$=149, $|E|$=16,538). RICH achieves Coverage 0.227 vs. LLM+DPP 0.591. Two densely cross-linked articles (Bay Area unionisation petitions, doc 1; NLRB complaints, doc 2) together capture 70.8% of the final score mass despite being 20% of the documents, forming a *resonance chamber* that starves minority documents. Evaluation questions probed minority-document details (worker perceptions; SEIU's role in funding Workers United), and the SEIU topic receives only 2.3% of mass and never reaches the selection window. This failure is identifiable at inference: when the top-2 documents hold >65% of mass in a 10-document event, minority-source coverage tends to degrade; the $(1-\alpha)$ document-prestige floor is a partial mitigation but insufficient when the dominant pair is exceptionally dense.

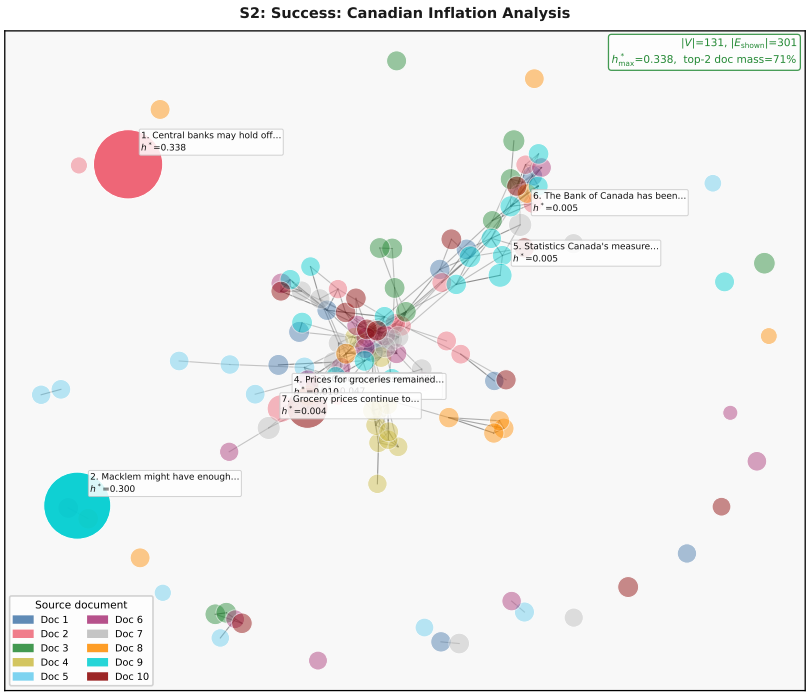

Figure 3: **Case S2: Canadian Inflation and Bank of Canada Rate Decision.** Topic graph after RecGNN convergence ($|V|$=131). The triangle of mutual reinforcement among the three articles that cite the same Statistics Canada release forms the dense central cluster. RICH Coverage 1.000; LLM+DPP 0.000.

Case F2: Ukraine War Situation Report (Analytical Question Mismatch). Ten articles on the battles for Bakhmut and Avdiivka ($|V|$=175, 10 docs). RICH achieves Coverage 0.000 vs. LLM+DPP 0.333. The RecGNN correctly surfaces well-connected tactical topics (holding Bakhmut, h^*=0.418; the noose around Avdiivka, h^*=0.257), but the questions probe geopolitical analysis (NATO as a trigger, strategic value of Bakhmut, de-escalation) that appears as sparsely connected peripheral nodes ($h^*<0.01$). RICH rewards *corroboration*, so it cannot distinguish "important but rare" from "important and redundant", a fundamental limitation of vertex-reinforced approaches. A post-hoc long-tail rescue step (Appendix M) or a stronger document-prestige floor could partly address this.

Lesson. Both failures reflect the same tension: RICH is optimized for *corroboration breadth*, whereas some questions probe *analytical depth* (perspectives unique to a minority of sources).

O Limitations and Broader Impact

Limitations. RICH has two structural failure modes identified through case analysis (Appendix N). *Document resonance bias*: when a subset of source documents is densely cross-linked, they absorb disproportionate score mass, dominating the ranked list even when not uniquely relevant. *Analytical question mismatch*: analytically important but sparsely corroborated topics (a single expert opinion or minority-source inference) are under-ranked by the vertex-reinforced walk, which rewards corroboration. Both are rooted in cosine-similarity edge construction; they are where one dense document pair dominates despite the per-source floor guarantee (Theorem 3), so they lie outside that guarantee rather than violate it. Future work will address them through multi-aspect intent expansion (Appendix L) and adversarial evaluation on rare-topic intents (Appendix M). More broadly, the empirical multi-document coverage is driven mainly by the diversity term rather than the hierarchical blend alone (Appendix J), and Theorem 2 guarantees existence and stability of the ranking, with the convergence rate measured directly rather than bounded in closed form.

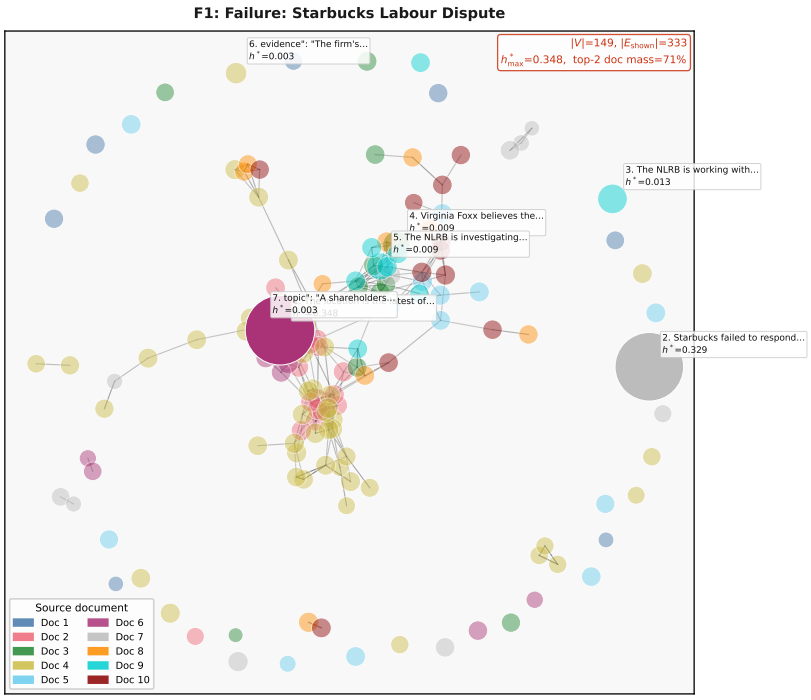

Figure 4: **Case F1: Starbucks Labour Dispute (Document Resonance Bias).** Doc 1 and Doc 2 together hold 70.8% of h^* mass, forming a resonance chamber; question-relevant minority-document topics appear as small peripheral nodes.

Broader impact. Structured multi-document summarization can help researchers, journalists, and knowledge workers efficiently surface diverse perspectives from large corpora, a net-positive contribution to information access. The same technology could be misused to produce engagement-optimized digests that selectively omit inconvenient sources or amplify a narrative. The explicit multi-source coverage objective and the stochastic-dominance guarantee are direct architectural mitigations against single-source bias, and the full evaluation protocol and hyperparameters (Appendix D) are provided so that coverage audits can be reproduced independently.

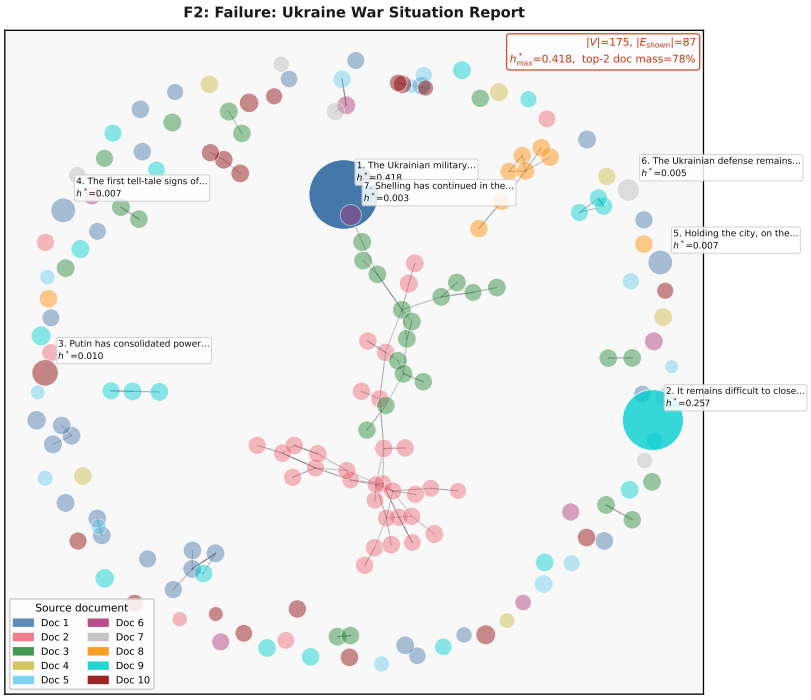

Figure 5: **Case F2: Ukraine War Situation Report (Analytical Question Mismatch).** The top-2 nodes are well-corroborated tactical reports; analytically important geopolitical topics appear as sparse peripheral nodes with $h^* < 0.01$. RICH Coverage 0.000; LLM+DPP 0.333.