HybridScholar: Personalized Paper Recommendation with Cross Attention-Gated Fusion and Verified Explanation Generation

Joytun Tonny[1], Sajidur Rahman[2], Saddam Mukta[3], Swakkhar Shatabda[4], Salekul Islam[5]
[1]United International University [2]Independent University Bangladesh [3]LUT University
[4]BRAC University [5]North South University

Abstract

The exponential growth of scientific literature has made personalized, factually grounded paper discovery increasingly difficult. Existing systems suffer from limited personalization, cold-start fragility, or hallucinated explanations. We introduce HYBRIDSCHOLAR, an integrated architecture combining a CrossAttention-gated fusion of collaborative filtering and RAG learning candidate-specific gate weights per candidate with a Chain-of-Verification (CoVe) pipeline for factually grounded explanations. Evaluated on the Scholar Inbox dataset, HYBRIDSCHOLAR substantially outperforms (NDCG@10 = 0.2991) strong baselines including LightGCN and Scholar-Inbox. Cold-start analysis shows the RAG fallback prevents hard failure for sparse users (NDCG@10 = 0.283), with cold NDCG rising monotonically as interactions grow. The CoVe pipeline reduces hallucination rate by 23.8%, confirmed by a human evaluation study showing CoVe-revised explanations are preferred by experts on factual accuracy (76.7% vs. 36.0%) and by end-users on helpfulness (78.8%, $p<0.0001$). Overall, HYBRIDSCHOLAR improves ranking quality, cold-start robustness, and factual reliability for large-scale scholarly research discovery, demonstrated on 15,881 users across 239,618 papers.

1 Introduction

Scientific publication volume has accelerated dramatically: the ACL Anthology alone grew ≈70% between 2019 and 2023 (Bollmann et al., 2023). Personalised recommender systems address discovery at scale by modelling user preferences from interaction histories, but their output is limited to ranked document lists, leaving a gap between paper discovery and the explanatory understanding true scientific engagement requires.

Large Language Models (LLMs) enable rapid natural-language interaction but frequently generate fabricated claims not grounded in retrieved sources (Ji et al., 2023), a particularly damaging failure mode given academic research's dependence on factual correctness and traceable provenance. Contemporary systems address these challenges in isolation: Scholar Inbox (Flicke et al., 2025) leverages interaction history but lacks integrated verification; RAG-centric tools such as GenGO Ultra (Takeshita et al., 2025) offer grounded QA but ignore individual researcher preferences.

To resolve these limitations, we propose HYBRIDSCHOLAR, an integrated architecture combining a CrossAttention-gated fusion of collaborative filtering (CF) and RAG for personalised ranking with a Chain-of-Verification (CoVe; Dhuliawala et al. 2024) pipeline for factually grounded explanations. Evaluated on the Scholar Inbox dataset (Flicke et al., 2025), HYBRIDSCHOLAR achieves NDCG@10=0.2991 and HR@10=0.5372, significantly outperforming all thirteen baselines. CoVe with a GPT-4o-Mini verifier reduces hallucination rate by 23.8% across 11,371 verified claims, confirmed by an independent NLI scorer.

Our contributions are as follows:

- CrossAttention fusion that learns candidate-specific CF and RAG gate weights, outperforming all fixed-weight and gating-based alternatives by up to +0.044 NDCG@10 (§5).
- A controlled verifier-capacity ablation (generator fixed) showing GPT-4o-Mini achieves 78% stronger effect size than a smaller verifier, within this comparison a driver of CoVe effectiveness (§5.4).
- Large-scale evaluation on 15,881 users with cold-start robustness analysis: the RAG fallback achieves competitive non-personalised performance and cold NDCG rises monotonically with interactions, alongside a counter-intuitive *negative* warm/cold gap under CF-

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 275–288
November 6-10, 2026 ©2026 Association for Computational Linguistics

Only traced to interaction noise for interdisciplinary readers (§5.3).

- Two complementary human evaluation studies: expert annotation (n=100 pairs, 3 annotators; κ=0.72) and blind end-user preference (n=72 trials, 12 participants) confirming automated faithfulness gains translate to human-perceived improvements in factual accuracy (+40.7 pp) and helpfulness (78.8% preference rate) (§5.5).

2 Related Work

Scientific Paper Recommendation. Content-based filtering (Wang et al., 2018; Patra et al., 2020) relies on paper attributes via TF-IDF (Jones, 1972) or dense representations (Cohan et al., 2020; Li et al., 2023a), but does not model evolving interests implicit in interaction histories (views, saves, highlights). Collaborative filtering (Murali et al., 2019; Sakib et al., 2020) models interactions but suffers from cold-start. Hybrid Dual-Encoder (Yi et al., 2019; Church et al., 2024; Sugiyama and Kan, 2013, 2015) or graph (Wang et al., 2024; Ostendorff et al., 2022) approaches address this trade-off; the Scholar Inbox dataset (Flicke et al., 2025) enables our large-scale evaluation. Our CrossAttention fusion extends Dual-Encoder designs by conditioning RAG weighting on the collaborative representation, producing a candidate-specific gate rather than a fixed blend. Closest in spirit is AgentDR (Yang et al., 2026) (LLM agents with traditional recommenders via tool-suitability selection), differing architecturally (agent-mediated selection vs. our differentiable gate) and in domain (e-commerce vs. scientific literature).

RAG for Scientific Text. Retrieval-Augmented Generation (Lewis et al., 2020) grounds LLM outputs in retrieved documents, reducing hallucination; GenGO Ultra (Takeshita et al., 2025) applies this with citation enforcement for scientific text generation. Standard RAG still suffers 30-40% hallucination rates (Maynez et al., 2020); attribution (Gao et al., 2023) and query rewriting (Ma et al., 2023) reduce but do not eliminate fabrication, motivating explicit claim verification.

Hallucination Mitigation. Chain-of-Verification (CoVe; Dhuliawala et al. 2024) generates verification questions, answers them against source documents, and revises unsupported claims, reducing hallucinations by 30-50% across long-form and list-based QA tasks. Related approaches include self-consistency sampling (Manakul et al., 2023), fact-checking with external tools (Peng et al., 2023), and attribution with verified excerpts (Menick et al., 2022). We extend CoVe to scholarly explanation grounding, replacing self-questioning with atomic claim decomposition and GPT-4o-Mini verification against retrieved abstracts, and provide a controlled ablation showing verifier capacity drives CoVe's faithfulness gains. "Rank, Don't Generate" (Kabongo et al., 2026) instead ranks pre-extracted statements rather than generating free text, eliminating hallucination by construction but requiring an extractable statement pool per item – less natural for our per-pair justifications. No prior system unifies personalised recommendation with verified QA; Table 1 summarises the gap.

Table 1: Comparison of recent related works with the proposed study. **CAF** = CrossAttention Fusion in Recommendation, **SRAG** = Semantic use of Retrieval-Augmented Generation, **HMR** = Hallucination Mitigation in Recommendation, **GEP** = Grounded Explanation of post-recommendation. ✓ = supported; × = not supported.

Paper	CAF	SRAG	HMR	GEP
Ostendorff et al. (2022)	×	×	×	×
Church et al. (2024)	×	×	×	×
Takeshita et al. (2025)	×	✓	×	✓
Dhuliawala et al. (2024)	×	×	✓	✓
Flicke et al. (2025)	×	×	×	×
Yang et al. (2026)	×	×	✓	×
Kabongo et al. (2026)	×	×	✓	✓
HybridScholar (Ours)	✓	✓	✓	✓

3 HybridScholar

We propose a hybrid scientific paper recommendation system that integrates collaborative filtering with RAG to provide personalized, explainable recommendations with hallucination mitigation.

3.1 Problem Formulation

Let $\mathcal{U} = \{u_1, \ldots, u_N\}$ be users and $\mathcal{P} = \{p_1, \ldots, p_M\}$ a paper corpus. For each user u_i, we define the **interaction history** as $\mathcal{H}_i = \{p_j \in \mathcal{P} \mid u_i \text{ has viewed, saved, or highlighted } p_j\}$, ordered chronologically by interaction timestamp. Given u_i, history $\mathcal{H}_i$, and an optional query q, the goal is to produce a ranked list of K recommendations with grounded explanations.

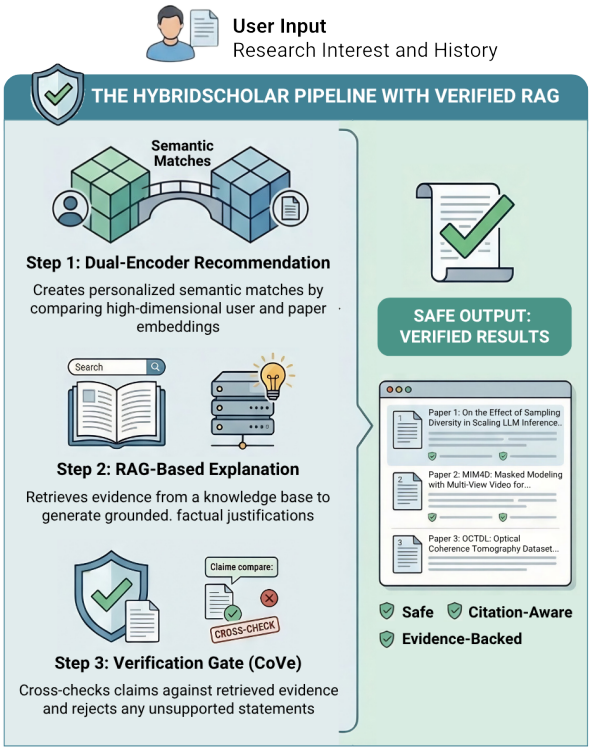


Figure 1: Overview of the HybridScholar system architecture. The system combines Dual-Encoder recommendation, RAG-based explanation, and verification to provide personalized, factually grounded paper recommendations.

3.2 System Architecture

Figure 1 overviews the system, addressing three limitations: (i) shallow personalisation in content-based systems modelling attributes but not implicit interaction history (Wang et al., 2018; Patra et al., 2020); (ii) popularity bias in graph-based CF (He et al., 2020b; Sugiyama and Kan, 2013); and (iii) hallucinated explanations in standard RAG (Maynez et al., 2020; Ji et al., 2023).

HybridScholar operates in three stages: (1) a Dual-Encoder collaborative filter generates candidate rankings from interaction history; (2) a RAG component retrieves relevant papers and generates grounded explanations; (3) a CrossAttention fusion module combines both signals into a final ranking, validated by a *CoVe* step against retrieved abstracts before output.

3.2.1 Dual-Encoder Collaborative Filter

We employ a Dual-Encoder comprising a Paper Encoder $f_{\text{paper}}(\cdot)$ and a User Encoder $f_{\text{user}}(\cdot)$ that project papers and users into a shared 128-dimensional latent space.

Paper Encoder. Each paper p_j is represented by a pre-computed GTE-Large embedding $\mathbf{e}_j \in \mathbb{R}^{256}$ (Li et al., 2023b), projected through two fully-connected layers with ReLU activations, batch normalisation, and dropout (0.2):

$$\mathbf{h}_j = \text{Lin}_{256\to128}(\text{drop}(\text{BN}(\text{ReLU}(\text{Lin}_{256\to512}(\mathbf{e}_j))))) \tag{1}$$

User Encoder. Given interaction history $\mathcal{H}_i$, we aggregate paper embeddings using attention-based pooling:

$$\begin{aligned} \mathbf{u}_i &= \sum_{p_j \in \mathcal{H}_i} \alpha_j \mathbf{e}_j, \\ \alpha_j &= \frac{\exp\left(\mathbf{w}^\top \tanh(\mathbf{W}\mathbf{e}_j)\right)}{\sum_{p_k \in \mathcal{H}_i} \exp(\mathbf{w}^\top \tanh(\mathbf{W}\mathbf{e}_k))} \end{aligned} \tag{2}$$

weighting the most predictive papers rather than treating all interactions equally.

Scoring. The CF score is cosine similarity in the shared latent space:

$$s_{\text{CF}}(u_i, p_j) = \frac{\mathbf{u}_i^\top \mathbf{h}_j}{\|\mathbf{u}_i\| \, \|\mathbf{h}_j\|} \tag{3}$$

Training. CF-Only and every fusion variant are trained end-to-end with an InfoNCE-style contrastive objective on the score s each model produces (s_{CF} for CF-Only; the fused score, Eq. 10, for fusion variants). The explicit-negative term scores the true positive pair against K sampled negatives:

$$\mathcal{L}_{\text{exp}} = -\log \frac{\exp(s^+/T)}{\exp(s^+/T) + \sum_{k=1}^{K} \exp(s_k^-/T)}, \tag{4}$$

with an in-batch term using other users' positive papers in the same batch as additional negatives:

$$\mathcal{L}_{\text{in-batch}} = -\frac{1}{B} \sum_{i=1}^{B} \log \frac{\exp(\mathbf{u}_i^\top \mathbf{h}_i/T)}{\sum_{k=1}^{B} \exp(\mathbf{u}_i^\top \mathbf{h}_k/T)}. \tag{5}$$

The final objective is $\mathcal{L}_{\text{CF}} = (1-\lambda)\,\mathcal{L}_{\text{exp}} + \lambda\,\mathcal{L}_{\text{in-batch}}$, with temperature T=0.1 and in-batch weight λ=0.3.

3.2.2 RAG System with Semantic Retrieval

The RAG component (Lewis et al., 2020) supports query- and history-based retrieval, synthesising a pseudo-query from recent interactions when no explicit query is available. Dense retrieval uses FAISS (Johnson et al., 2017) over L2-normalised GTE-Large embeddings:

$$s_{\text{RAG}}(q, p_j) = \cos(\mathbf{e}_q, \mathbf{e}_j) \tag{6}$$

The top-5 retrieved abstracts are passed as grounding documents to Llama-3.1-8b-Instant (DeepSeek Chat as fallback) to generate concise 1-2 sentence contextual justifications for each recommendation.

3.2.3 CrossAttention Fusion

Prior fusion approaches apply a fixed linear blend $s = \alpha\, s_{\mathrm{CF}} + \beta\, s_{\mathrm{RAG}}$, unable to adapt per-pair compatibility. We instead propose a **CrossAttention** gating mechanism conditioning the fusion weight on the learned CF representations before combining the two scores.

Cross-Modal Attention on Latent Representations. Rather than operating on the scalar scores s_{CF} and s_{RAG} directly, the CrossAttention module takes the d-dimensional latent embeddings produced by the Dual-Encoder (Eqs. 1-2) as input. Attending over the single pooled user vector $\mathbf{u}_i$ would make the key/value sequence length 1, so the softmax below would be identically 1 and the resulting gate would depend only on the user, not the candidate. We instead let the candidate embedding attend over the user's *individual* history-item embeddings: let $\mathbf{H}_i = [\mathbf{h}_{i,1}; \ldots; \mathbf{h}_{i,|\mathcal{H}_i|}] \in \mathbb{R}^{|\mathcal{H}_i| \times d}$ be the matrix of per-paper embeddings (Eq. 1) for the papers in u_i's interaction history $\mathcal{H}_i$, using the same padding mask as the attention pooling in Eq. 2. We use the candidate embedding $\mathbf{h}_j \in \mathbb{R}^d$ as query and $\mathbf{H}_i$ as both key and value in a multi-head cross-attention layer:

$$\mathbf{a}_j = \mathrm{MultiHead}(\mathbf{h}_j,\ \mathbf{H}_i,\ \mathbf{H}_i) \in \mathbb{R}^d, \quad (7)$$

with n_h=2 attention heads and d_k=d/n_h=64. Internally, each head ℓ computes:

$$\mathrm{head}_\ell = \mathrm{softmax}\left(\frac{\mathbf{h}_j \mathbf{W}_Q^\ell (\mathbf{H}_i \mathbf{W}_K^\ell)^\top}{\sqrt{d_k}}\right) \mathbf{H}_i \mathbf{W}_V^\ell, \quad (8)$$

where $\mathbf{W}_Q^\ell, \mathbf{W}_K^\ell, \mathbf{W}_V^\ell \in \mathbb{R}^{d \times d_k}$ are per-head learned projections, and heads are concatenated and linearly projected back to $\mathbb{R}^d$. Because the key/value sequence now has length $|\mathcal{H}_i|$ rather than 1, the softmax is non-degenerate whenever $|\mathcal{H}_i| \geq 2$[1], so different candidates attend differently over the same user's history, producing a genuinely candidate-specific compatibility signal $\mathbf{a}_j$ (empirically verified in Appendix A.3).

[1]For users with a single history item ($|\mathcal{H}_i|$=1), the attention weight is trivially 1 and g_j reduces to that one paper's compatibility with the candidate – still candidate-specific, unlike the original pooled-vector formulation.

Candidate-Specific Gating. The attended representation is passed through a learned gate to produce a candidate-specific scalar fusion weight:

$$g_j = \sigma\big(\mathbf{w}^\top \mathbf{a}_j + b\big) \in (0, 1), \quad (9)$$

where σ is the sigmoid function and $\mathbf{w} \in \mathbb{R}^d$, $b \in \mathbb{R}$ are learnable parameters. When cross-attention finds strong agreement between paper and user representations, g_j increases, up-weighting RAG; when agreement is weak, g_j decreases, relying more on CF.

Hybrid Ranking Score. The final ranking score for candidate p_j is the gate-weighted combination of the CF and RAG scores:

$$\begin{aligned} s(u_i, q, p_j) = (1 - g_j)\ s_{\mathrm{CF}}(u_i, p_j) \\ +\ g_j\ s_{\mathrm{RAG}}(q, p_j), \end{aligned} \quad (10)$$

where s_{CF} and s_{RAG} are the scalar scores from Eqs. (3) and (6), combined with the candidate-specific gate g_j rather than the fixed $\alpha\, s_{\mathrm{CF}} + \beta\, s_{\mathrm{RAG}}$ blend.

3.2.4 Hallucination Mitigation via CoVe

To improve factual reliability, we adopt Chain-of-Verification (CoVe; Dhuliawala et al. 2024), which extracts individual claims, verifies each against retrieved abstracts using a zero-temperature LLM, and revises unsupported claims. Baseline and CoVe conditions use identical generation temperature (T=0.3) to eliminate confounding.

We evaluate faithfulness on n=500 matched explanation pairs using two complementary scorers: *NLI-based entailment* (DeBERTa-v3-large; He et al. 2020a) and *MiniCheck* (Tang et al., 2024), measuring mean sentence-level groundedness against retrieved abstracts. *ALCE* (Gao et al., 2023) requires gold citation annotations unavailable in our abstract-only setting; *FactScore* (Min et al., 2023) requires entity-level knowledge bases; *RAGAS* (Es et al., 2024) relies on an LLM judge whose cost scales with explanation volume. MiniCheck and NLI-DeBERTa need neither and are purpose-built for claim-level grounding at scale. Statistical significance is assessed via paired t-test and Wilcoxon signed-rank test, with Cohen's d and 95% bootstrap CIs reported.

4 Experimental Setup

4.1 Dataset and Preprocessing

We use the publicly released Scholar Inbox dataset (Flicke et al., 2025)[2], combining its three interaction signals – (1) *explicit ratings*, (2) *implicit interactions* (views, saves), and (3) *abstract highlighting* – with arXiv metadata into a corpus of *239,618 papers*, each encoded with GTE-Large (Li et al., 2023b) (1024-d, PCA-compressed to 256-d retaining 95% variance).

The combined dataset contains *15,881 users* and *1,003,241 interactions*. Splitting is performed *per user* rather than via a single global temporal cutoff: each user's history is independently partitioned, chronologically, into training, validation, and test segments (nominally 70/10/20), preventing future-interaction leakage. Validation is used solely for early stopping; all reported results use the held-out *test* partition (194,404 interactions). Each test instance is evaluated by ranking the target against the full 239,618-paper corpus, avoiding the sampled-negative evaluation bias documented by (Krichene and Rendle, 2020). Users with $|\mathcal{H}_i| < 5$ training interactions are designated cold-start. Dataset statistics are in Table 2.

Table 2: Dataset statistics for the arXiv CS corpus used in all experiments. Splitting is performed per user, not via a single global cutoff (§4), yielding 714,055 / 94,782 / 194,404 train/validation/test interactions across 15,881 users; all reported results use the held-out test partition.

Statistic	Count
Paper Corpus	
Total papers	239,618
Embedding dimension	1024 → 256 (GTE-Large → PCA)
Variance retained	95%
User Interactions	
Total users	15,881
Training interactions	714,055
Validation interactions	94,782
Test interactions[‡]	194,404
Total interactions	1,003,241
Evaluation	
Ranking pool	full corpus (239,618)
Cold-start threshold	<5 interactions
Evaluation instances	194,404

[‡]Used for all reported results in this paper.

[2] https://github.com/avg-dev/scholar_inbox_datasets

4.2 Baselines, Metrics, and Significance

We compare against thirteen systems: *Popularity*, *BM25* (Robertson et al., 2009), and *TF-IDF* (Salton and Buckley, 1988) (non-personalised IR baselines (Sugiyama and Kan, 2013, 2015)); *LightGCN* (He et al., 2020b) (graph-based CF); *Scholar-Inbox* (Flicke et al., 2025) (reimplemented on our corpus with identical embeddings); *Item-kNN* (max cosine similarity to any single history item, zero trainable parameters), *BPR-MF* (ID-embedding matrix factorisation with the standard BPR pairwise loss, deliberately not built on the content embeddings used elsewhere), and *Sequential-GRU* (a GRU over the chronologically-ordered history, final hidden state scored against the candidate); and five fusion ablations — *CF-Only* (α=1), *LLM-Direct* (β=1), *SimpleFusion* (fixed α=0.7, β=0.3), *MultiHeadFusion* (self-attention over CF+RAG), and *NeuralGatingFusion* (learned sigmoid gate) — each isolating a specific design decision in our CrossAttention architecture. CF-Only is itself a two-tower model (separate Paper/User Encoders, Eqs. 1-2); LightGCN provides a graph-based CF comparison.

We report *NDCG@10* (primary) (Järvelin and Kekäläinen, 2002), *HR@10*, and *MRR* (Voorhees et al., 1999). Recall@10 is omitted: under our single-relevant-item protocol it is mechanically identical to HR@10. P@5/P@10 are likewise mechanically capped (0.2/0.1) and reported for completeness only, not comparison. Significance is assessed via bootstrap resampling (n=1,000, 95% CI) and Wilcoxon signed-rank tests ($^{***}p<0.001$, $^{**}p<0.01$, $^{*}p<0.05$). NDCG@10/HR@10 are computed by ranking each target against the full 239,618-paper corpus rather than a sampled candidate pool, avoiding the overstatement risk that sampled-negative evaluation carries (Krichene and Rendle, 2020).

GenGO Ultra (Takeshita et al., 2025) is excluded from Table 3: it is a text-generation system with no personalization component and cannot produce per-user rankings; LLM-Direct (RAG-only) already captures the equivalent RAG-based scoring signal under our protocol. LLM-based rankers (e.g. (Bao et al., 2025)) are also not compared: (1) serialising a user's full history into a prompt is impractical at our scale (up to 50 items × 15,881 users × 239,618 candidates per instance) without confounding subsampling; (2) no open-source LLM ranker fits the Scholar Inbox schema (ratings, views, saves,

highlights), requiring a new training pipeline rather than a drop-in baseline; (3) LLM-Direct (β=1) already isolates a pure LLM-mediated signal, well below CrossAttention (Table 3). We leave a rigorous LLM-reranker comparison to future work (§7).

4.3 Implementation Details

Full hyperparameter configurations for all system components are reported in Appendix A.1. The Dual-Encoder is trained with the InfoNCE-style contrastive objective of Eqs. 4-5 (T=0.1, in-batch weight λ=0.3) and Adam (Kingma and Ba, 2014) (lr=10^{-3}). K=16 negatives per positive; 8 semantic hard negatives (FAISS top-50 neighbours excluding H_i) and 8 uniform random negatives. All fusion variants share an identical Dual-Encoder backbone, training data, optimiser, learning rate, and epoch budget. The gate is not trained separately: s_{CF}, s_{RAG}, and gate parameters are optimised jointly via the same objective applied to the fused score (Eq. 10); both are cosine similarities between L2-normalised embeddings, with s_{RAG} computed live per candidate rather than cached (§3.2.3). The CrossAttention fusion uses d_k=64 projections and a 2-layer ReLU MLP (297,602 parameters total). Explanations are generated at T=0.3 for both RAG-only and CoVe conditions. Faithfulness is scored on n=500 pairs by two independent automated scorers: DeBERTa-v3-large NLI (He et al., 2020a) and MiniCheck-Flan-T5-Large (Tang et al., 2024). Full latency and cost breakdown for a single user session is reported in Table 8 (Appendix A).

5 Results and Analysis

5.1 Main Results

Table 3 reports results for all 14 systems on 194,404 held-out test instances across 15,881 users. All comparisons against CrossAttention are significant at p<0.001 (bootstrap 95% CI, n=1,000 resamples; Wilcoxon signed-rank).

CrossAttention substantially outperforms all non-personalised baselines (+0.253 NDCG@10 over Popularity), confirming interaction-based personalisation is essential for scholarly recommendation (Sugiyama and Kan, 2013, 2015). Against the strongest single-component methods, CrossAttention achieves **+0.065** NDCG@10 over CF-Only and **+0.068** over LLM-Direct, showing cross-signal *interaction*, not mere score combination, drives performance; it also surpasses Scholar-Inbox (Flicke et al., 2025) by +0.070 and LightGCN by +0.210 (both p<0.001). Sequential-GRU (0.2218) and Item-kNN (0.1928) fall short of every fusion variant, while BPR-MF (0.0675) performs comparably to LightGCN, indicating ID-only or single-item signals are insufficient without richer collaborative/semantic fusion. The fusion ablations reveal a clear monotone ordering: SimpleFusion (0.2554) < MultiHeadFusion (0.2591) < NeuralGatingFusion (0.2713) < CrossAttention (0.2991), the +0.028 gap over NeuralGatingFusion confirming that conditioning RAG weighting on the CF representation, not treating both signals symmetrically, is the key architectural advantage.

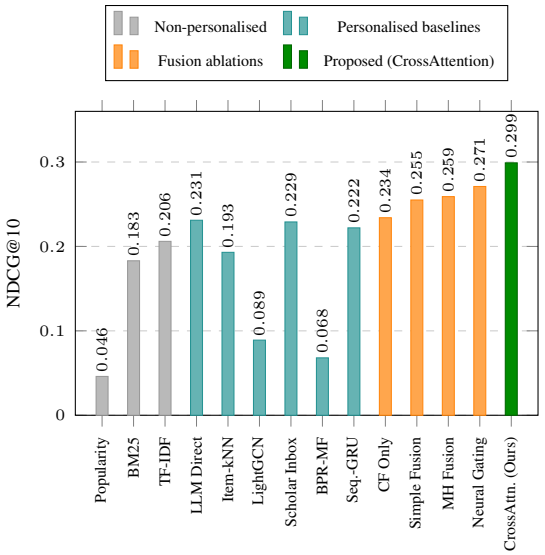


Figure 2: NDCG@10 for all 14 systems. Non-personalised baselines (grey), personalised baselines (teal), fusion ablations (orange), and proposed CrossAttention model (green). Scores computed with the corrected candidate-specific gate (§3.2.3). All differences vs. CrossAttention are significant at p<0.001 (Wilcoxon).

5.2 Ablation Study

Figure 3 (Appendix A.2) visualises this ordering across NDCG@10, HR@10, and MRR. NeuralGatingFusion also outperforms MultiHeadFusion on MRR (0.2763 vs. 0.2326), suggesting a learned per-user gate helps first-rank placement, not just relevance selection; CrossAttention's +0.028 MRR gain over NeuralGatingFusion (p<0.001) follows the same pattern as its NDCG@10 advantage.

5.3 Cold-Start Analysis

Table 4 and Table 9 (Appendix A.4) extend this across $\tau \in \{5, 10, 15, 20\}$ for CrossAttention: the

Table 3: Table 3: Performance comparison on 15,881 users (194,404 held-out test instances, ranked against the full 239,618-paper corpus), computed with the candidate-specific CrossAttention gate and live per-candidate s_{RAG} 3.2.3. Best in **bold**, second-best underlined. R@10 is omitted as it is mechanically identical to HR@10 under our single-relevant-item protocol. 95% CI on NDCG@10 via bootstrap (n=1,000). $^{***}p<0.001$ vs. proposed model (Wilcoxon signed-rank). †Reimplemented on our corpus.

Model	P@5	P@10	NDCG@10	95% CI	HR@10	MRR
Non-Personalised						
Popularity	0.0109	0.0091	0.0463	[0.0451, 0.0475]	0.0906	0.0436 ***
BM25	0.0446	0.0312	0.1827	[0.1803, 0.1851]	0.3122	0.1578 ***
TF-IDF	0.0503	0.0354	0.2060	[0.2035, 0.2085]	0.3540	0.1763 ***
Personalised Baselines						
LLM-Direct (RAG)	0.0572	0.0394	0.2308	[0.2282, 0.2334]	0.3943	0.1965 ***
Item-kNN	0.0472	0.0351	0.1928	[0.1904, 0.1952]	0.3509	0.1628 ***
LightGCN	0.0216	0.0160	0.0894	[0.0878, 0.0910]	0.1599	0.0791 ***
Scholar-Inbox†	0.0567	0.0392	0.2288	[0.2262, 0.2314]	0.3920	0.1933 ***
BPR-MF	0.0163	0.0133	0.0675	[0.0661, 0.0689]	0.1334	0.0601 ***
Sequential-GRU	0.0548	0.0390	0.2218	[0.2192, 0.2244]	0.3901	0.1870 ***
Fusion Ablations						
CF-Only	0.0576	0.0408	0.2342	[0.2316, 0.2368]	0.4076	0.1979 ***
SimpleFusion	0.0626	0.0439	0.2554	[0.2527, 0.2581]	0.4385	0.2152 ***
MultiHeadFusion	0.0673	0.0447	0.2591	[0.2564, 0.2618]	0.4466	0.2326 ***
NeuralGatingFusion	0.0697	0.0461	0.2713	[0.2685, 0.2741]	0.4812	0.2763 ***
Proposed						
CrossAttention	**0.0730**	**0.0497**	**0.2991**	[0.2962, 0.3020]	**0.5372**	**0.3046**

Table 4: Cold-start vs. warm-start NDCG@10 at τ=5, computed with the corrected candidate-specific CrossAttention gate and live per-candidate s_{RAG} (§3.2.3); this applies to all RAG-consuming rows (CrossAttention, NeuralGating, SimpleFusion). Gap = warm − cold. $^{***}p<0.001$ (Wilcoxon). CF-Only uses no RAG signal.

Model	Cold	Hit_c	Warm	Hit_w	Gap	d
CrossAttention	0.2834	0.4492	**0.4769**	**0.4968**	+0.1935	1.066^{***}
NeuralGating	0.2051	0.3449	0.4269	0.4665	+0.2218	0.984^{***}
SimpleFusion	0.2902	0.4620	0.3681	0.3933	+0.0779	0.651^{***}
CF-Only	**0.2915**	**0.4788**	0.2451	0.2845	−0.0464	-0.055^{***}

standardised gap (Cohen's d) narrows monotonically with history even as the raw NDCG@10 gap widens slightly, since within-group variance grows faster than the mean difference as the warm bucket's history requirement tightens. Cold-start scores remain competitive with full-data baselines (Table 3), confirming the RAG fallback is not degenerate.

A stricter *true new-user* check - history forcibly emptied, isolating the RAG-only pathway from residual personalisation among nominally cold-start users with 1–4 interactions - gives NDCG@10=0.0843, HR@10=0.1996 (n=500), well below the full personalised result (0.2991, Table 3) but still well above chance.

CF-Only alone shows a *negative* warm/cold gap (warm-start users score worse), attributable to interaction noise degrading representations for interdisciplinary readers (cold-start users rely on popularity alone). Since field metadata covers only 3.7% of papers (8,817/239,618), we instead correlated history diversity (1−mean pairwise cosine similarity among history embeddings) with CF-Only NDCG@10 over 1,500 warm-start users: negative and significant (Pearson $r=-0.36$, $p<10^{-46}$). CrossAttention avoids this by cross-attending CF with RAG evidence, motivating RAG as a ranking stabiliser, not only an explanation generator.

5.4 Explanation Faithfulness

Table 5: Explanation faithfulness on n=500 pairs (T=0.3, GPT-4o-Mini verifier). MiniCheck is the primary scorer (Tang et al., 2024); NLI is cross-validation (He et al., 2020a). $^{***}p<0.001$ (paired t-test + Wilcoxon).

Scorer	System	Faith↑	Δ	d	Sig.
MiniCheck (primary)	RAG only	0.6520	—	0.469	***
	+CoVe (GPT-4o-Mini)	**0.7349**	+0.083		
NLI (cross-val.)	RAG only	0.1981	—	0.375	***
	+CoVe (GPT-4o-Mini)	**0.2747**	+0.077		

Table 5 reports faithfulness on n=500 matched pairs (generator fixed; all improvement attributable to CoVe): both scorers show significant hallucination reduction ($p<0.001$; d=0.469, d=0.375). MiniCheck scores atomic claims directly; NLI operates at sentence level and penalises valid

paraphrase, yielding lower absolute values by design (Tang et al., 2024). The low inter-scorer Spearman ρ=0.138 reflects this granularity difference, not disagreement; direction concordance of 62.4% (vs. 50% chance; χ^2=15.2, p<0.001) confirms convergent validity.

Verifier capacity ablation. Table 6 isolates verifier choice with the generator held constant. Upgrading from Llama-3.1-8b to GPT-4o-Mini yields substantial gains in both MiniCheck and NLI effect sizes, with revision rate rising accordingly, driven by GPT-4o-Mini's atomic decomposition surfacing unsupported specifics (fabricated metrics, nonexistent datasets, overclaimed comparisons) that smaller verifiers accept. This isolates one controlled comparison, not a full sweep; verifier capacity rather than generator capacity drives the gain here, though a broader sweep would be needed to generalise.

Table 6: Verifier capacity ablation on n=500 pairs. Generator (Llama-3.1-8b, T=0.3) held constant.

Verifier	MC Faith	d_{MC}	d_{NLI}	Revised
None (RAG only)	0.652	—	—	—
CoVe + Llama-3.1-8b	0.688	0.264	0.109	50.0%
CoVe + GPT-4o-Mini	**0.735**	**0.469**	**0.375**	83.6%

5.5 Human Evaluation

To validate that automated faithfulness gains translate to human-perceived quality, we conducted two studies comparing CoVe-revised against RAG-only explanations (Table 10, Appendix A.5).

Study 1: Expert Annotation. Three CS researchers independently annotated 100 blind explanation pairs stratified by MiniCheck score quartile across *factual accuracy* (binary), *usefulness*, and *naturalness* (5-point Likert), with counterbalanced presentation order.

Study 2: Blind Pairwise Preference. Twelve CS graduate students each evaluated six blind recommendation scenarios (≈8 min) as counterbalanced, unlabelled A/B pairs, judging (Q1) helpfulness and (Q2) factual reliability in forced-choice trials.

CoVe outperforms RAG-only on every dimension in both studies (Table 10). Study 1 agreement is substantial for factual accuracy and moderate for usefulness/naturalness, with the directional finding consistent across all three annotators. Study 2 preferences on both helpfulness and trustworthiness are significant with medium-to-large effect sizes (sign test p<0.0001). The two studies converge independently: automated hallucination reduction translates to human-perceived improvements in factual reliability and usefulness in a real-world research-discovery context.

5.6 Qualitative Analysis

Table 11 (Appendix A.6) summarises aggregate CoVe statistics across n=500 pairs; representative examples across all four behavioural categories (large/moderate/no/negative revision), stratified by baseline hallucination score (low: H_{base}<0.4; mid: 0.4–0.7; high: H_{base}>0.7), are in the same appendix. Faithfulness gains also occur in no-revision cases (e.g., H_{base}=0.976→H_{CoVe}=0.293), indicating the verification pass conditions the generator toward more conservative, abstract-grounded language even without explicit revision. Negative-revision cases, where atomic decomposition overcorrects already-faithful explanations, reflect a known limitation of salience-agnostic verifiers.

6 Conclusion

HYBRIDSCHOLAR integrates CrossAttention-gated CF and RAG fusion with a Chain-of-Verification pipeline, achieving NDCG@10=0.2991 on the Scholar Inbox dataset (+0.028 over the next best fusion variant, NeuralGatingFusion, p<0.001); ablation confirms that conditioning RAG weighting on the collaborative representation per candidate rather than fixed blending is the decisive advantage; the negative CF-Only warm/cold gap (§5.3) confirms RAG stabilises ranking across the interaction spectrum, not merely explanations. CoVe reduces hallucination rate by 23.8% (MiniCheck: 0.348→0.265, d=0.469, p<0.001), corroborated by human evaluation: experts prefer CoVe-revised explanations on factual accuracy at 76.7% vs.36.0% (+40.7 pp, κ=0.72) and end-users on helpfulness at 78.8% (p<0.0001, h=0.614). We release code and evaluation pipeline for reproducibility and further scholarly-recommendation research.

7 Limitations

Cold-start performance gap. Cold-start users ($|\mathcal{H}_i|$<5) achieve NDCG@10 = 0.283 versus 0.477 for warm-start users (d = 1.07). This gap reflects a transient data-efficiency challenge rather than an architectural failure: the RAG fallback prevents the hard degradation of pure CF, and cold

NDCG rises monotonically ($0.283 \rightarrow 0.344$, Table 9) as interactions grow, while the standardised gap (d) narrows over the same range ($1.07 \rightarrow 0.71$). Even under a stricter zero-history condition (history forcibly emptied rather than merely <5 interactions), NDCG@10 remains at 0.084, well above a non-personalised floor confirming the RAG fallback is not degenerate. Future work should explore meta-learning or few-shot personalisation (Schein et al., 2002) to accelerate warm-up beyond the five-interaction threshold.

LLM-based reranker baseline. Table 3 does not include an LLM-based reranker (e.g., (Bao et al., 2025)); §4 covers why serialising a full interaction history into a prompt at our scale, and the lack of an off-the-shelf ranker fit to the Scholar Inbox schema, rule out a drop-in comparison this cycle. LLM-Direct (RAG-only) gives a lower-bound sense of what a pure LLM-mediated signal contributes, but it isn't a substitute: a properly trained reranker could plausibly close some of the $+0.068$ NDCG@10 gap CrossAttention holds over it (§5.1). Of everything left out of the baseline set, this is the comparison we'd prioritise next.

Abstract-only retrieval. The RAG component grounds claims exclusively in paper abstracts. Full-text retrieval would enable finer-grained verification and likely reduce the 9.8% claim flagging rate, but substantially increases storage and inference cost at the scale of 239,618 papers. A selective strategy that triggers full-text retrieval only for low-confidence abstracts is a practical middle ground worth exploring.

Human evaluation scale. Our human studies involve 3 expert annotators (100 blind pairs) and 12 end-users (72 decisive trials). While results are statistically significant with medium-to-large effect sizes, a larger-scale study with diverse annotator backgrounds would further strengthen cross-domain generalisability.

Single-domain corpus. All experiments use arXiv computer science papers. Recommendation and explanation quality on biomedical, physics, or social science literature may differ due to variation in abstract informativeness, citation density, and terminological specificity. Domain transfer experiments using the Scholar Inbox (Flicke et al., 2025) dataset's full cross-domain breadth are a natural and important extension.

LLM API dependency. Explanation generation relies on Llama-3.1-8b-instant via the Groq API, introducing both latency (≈2-3 s per explanation) and a dependency on third-party infrastructure. Explanations are generated on-demand for the top-K recommendations per user session, not across the full 239,618-paper corpus; the 2–3s latency therefore applies to a small number of outputs per interaction. Distilling the generation and verification capability into a smaller, locally deployable model would improve reproducibility and reduce operational cost in resource-constrained settings.

References

Keqin Bao, Jizhi Zhang, Wenjie Wang, Yang Zhang, Zhengyi Yang, Yanchen Luo, Chong Chen, Fuli Feng, and Qi Tian. 2025. A bi-step grounding paradigm for large language models in recommendation systems. *ACM Transactions on Recommender Systems*, 3(4):1–27.

Marcel Bollmann, Nathan Schneider, Arne Köhn, and Matt Post. 2023. Two decades of the ACL anthology: Development, impact, and open challenges. In *Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023)*, pages 83–94, Singapore, Singapore. Empirical Methods in Natural Language Processing.

Kenneth Church, Omar Alonso, Peter Vickers, Jiameng Sun, Abteen Ebrahimi, and Raman Chandrasekar. 2024. Academic article recommendation using multiple perspectives. In *arXiv preprint*. ArXiv:2407.05836.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. 2020. SPECTER: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2270–2282, Online.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2024. Chain-of-verification reduces hallucination in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3563–3578, Bangkok, Thailand. Association for Computational Linguistics.

Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. RAGAs: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158, St. Julians, Malta. Association for Computational Linguistics.

Markus Flicke, Glenn Angrabeit, Madhav Iyengar, Vitalii Protsenko, Illia Shakun, Jovan Cicvaric, Bora

Kargi, Haoyu He, Lukas Schuler, Lewin Scholz, Kavyanjali Agnihotri, Yong Cao, and Andreas Geiger. 2025. Scholar inbox: Personalized paper recommendations for scientists. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 307–317, Vienna, Austria. Association for Computational Linguistics.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488, Singapore. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020a. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. 2020b. Lightgcn: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 639–648.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. In *ACM Computing Surveys*, volume 55, pages 1–38.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2017. Billion-scale similarity search with gpus. *Preprint*, arXiv:1702.08734.

Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. In *Journal of Documentation*, volume 28, pages 11–21.

Ben Kabongo, Arthur Satouf, and Vincent Guigue. 2026. Rank, don't generate: Statement-level ranking for explainable recommendation. *arXiv preprint arXiv:2604.03724*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Walid Krichene and Steffen Rendle. 2020. On sampled metrics for item recommendation. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1748–1757.

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

Kenneth Li, Oam Patel, Fernanda Viegas, Hanspeter Pfister, and Martin Wattenberg. 2023a. Inference-time intervention: Eliciting truthful answers from a language model. In *arXiv preprint*. ArXiv:2306.03341.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023b. Towards general text embeddings with multi-stage contrastive learning. In *arXiv preprint*. ArXiv:2308.03281.

Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. 2023. Query rewriting in retrieval-augmented large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5303–5315, Singapore. Association for Computational Linguistics.

Potsawee Manakul, Adian Liusie, and Mark JF Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *arXiv preprint*. ArXiv:2303.08896.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *arXiv preprint*. ArXiv:2005.00661.

Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, and 1 others. 2022. Teaching language models to support answers with verified quotes. In *arXiv preprint*. ArXiv:2203.11147.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FactScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *arXiv preprint*. ArXiv:2305.14251.

M Viswa Murali, T G Vishnu, and Nancy Victor. 2019. A collaborative filtering based recommender system for suggesting new trends in any domain of research. In *2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS)*, pages 550–553.

Malte Ostendorff, Nils Rethmeier, Isabelle Augenstein, Bela Gipp, and Georg Rehm. 2022. Neighborhood contrastive learning for scientific document representations with citation embeddings. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11670–11688, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Braja Gopal Patra, Vahed Maroufy, Babak Soltanalizadeh, Nan Deng, W. Jim Zheng, Kirk Roberts, and

Hulin Wu. 2020. A content-based literature recommendation system for datasets to improve data reusability - a case study on gene expression omnibus (geo) datasets. In *Journal of Biomedical Informatics*, volume 104, page 103399.

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, and 1 others. 2023. Check your facts and try again: Improving large language models with external knowledge and automated feedback. In *arXiv preprint*. ArXiv:2302.12813.

Stephen Robertson, Hugo Zaragoza, and 1 others. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Nazmus Sakib, Rodina Binti Ahmad, and Khalid Haruna. 2020. A collaborative approach toward scientific paper recommendation using citation context. In *IEEE Access*, volume 8, pages 51246–51255.

Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523.

Andrew I Schein, Alexandrin Popescul, Lyle H Ungar, and David M Pennock. 2002. Methods and metrics for cold-start recommendations. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 253–260.

Kazunari Sugiyama and Min-Yen Kan. 2013. Exploiting potential citation papers in scholarly paper recommendation. In *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries*, pages 153–162.

Kazunari Sugiyama and Min-Yen Kan. 2015. A comprehensive evaluation of scholarly paper recommendation using potential citation papers. *International Journal on Digital Libraries*, 16(2):91–109.

Sotaro Takeshita, Tornike Tsereteli, and Simone Paolo Ponzetto. 2025. GenGO ultra: an LLM-powered ACL paper explorer. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 242–251, Vienna, Austria. Association for Computational Linguistics.

Liyan Tang, Philippe Laban, and Greg Durrett. 2024. Minicheck: Efficient fact-checking of llms on grounding documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8818–8847.

Ellen M Voorhees and 1 others. 1999. The trec-8 question answering track report. In *Trec*, volume 99, pages 77–82.

Donghui Wang, Yanchun Liang, Dong Xu, Xiaoyue Feng, and Renchu Guan. 2018. A content-based recommender system for computer science publications. In *Knowledge-Based Systems*, volume 157, pages 1–9.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Improving text embeddings with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11897–11916, Bangkok, Thailand. Association for Computational Linguistics.

Mingdai Yang, Nurendra Choudhary, Jiangshu Du, Edward W Huang, Philip Yu, Karthik Subbian, and Danai Koutra. 2026. Agentdr: Dynamic recommendation with implicit item-item relations via llm-based agents. In *Proceedings of the ACM Web Conference 2026*, page 6230–6241. ACM.

Xinyang Yi, Ji Yang, Lichan Hong, Derek Zhiyuan Cheng, Lukasz Heldt, Aditee Kumthekar, Zhe Zhao, Li Wei, and Ed Chi. 2019. Sampling-bias-corrected neural modeling for large corpus item recommendations. In *Proceedings of the 13th ACM Conference on Recommender Systems*, pages 269–277, New York, NY, USA. Association for Computing Machinery.

A Appendix

A.1 Implementation Details

Table 7 reports the full hyperparameter configuration for all system components. All experiments were conducted on a single NVIDIA GPU with the Scholar Inbox dataset (Flicke et al., 2025).

A.2 Fusion Ablation Comparison

Figure 3 plots the four fusion variants of Table 3 (Fusion Ablations block) across NDCG@10, HR@10, and MRR, visualising the monotone ordering discussed in §5.

A.3 Gate Diagnostics

We verify empirically that the §3.2.3 (Eqs. 7-8) yields a genuinely candidate-specific gate rather than the degenerate, user-only formulation it replaces. For a sample of validation users, we score all candidates for each user in a single forward pass and record the resulting gate g_j alongside $s_{\mathrm{CF}}(u_i, p_j)$ and $s_{\mathrm{RAG}}(q, p_j)$ for every candidate. Under the original pooled-vector formulation, g_j is constant within a user by construction (gate std $= 0$, since the softmax cannot depend on the candidate); under the fixed model we observe a mean *within-user* gate standard deviation of **0.2253** (non-zero, confirming candidate-specific variation). The gate correlates negatively with both component scores:

Table 7: Hyperparameter and implementation details for HYBRIDSCHOLAR. All fusion variants share the same Dual-Encoder backbone and training configuration.

Component	Configuration
Dual-Encoder	
Input embedding dim.	256 (GTE-Large + PCA)
Hidden layers	[512, 256]
Output dimension	128
Dropout	0.2
User encoder	Attention-based pooling
Training	
Optimiser	Adam (Kingma and Ba, 2014)
Learning rate	1×10^{-3}
Batch size	16
Epochs	20
Loss function	InfoNCE + in-batch term
Temperature T	0.1
In-batch weight λ	0.3
CrossAttention Fusion	
Attention dim. d_k	64
MLP layers	2
Total parameters	297,602
Ranking pool	full corpus (239,618)
RAG System	
Embedding model	GTE-Large (thenlper)
Original dimension	1,024
Reduced dimension	256 (PCA, 95% var.)
Retrieval index	FAISS Flat L2
LLM (generation)	Llama-3.1-8b-Instant
LLM (verification)	GPT-4o-Mini
Temperature	0.3 (generation) 0.0 (verification)
CoVe Pipeline	
Evaluation pairs	500
Primary scorer	MiniCheck (flan-t5-large)
Cross-validation scorer	NLI (DeBERTa-v3-large)

$r = -0.73$ with s_{CF} and $r = -0.45$ with s_{RAG}, consistent with the gate suppressing whichever signal is comparatively weaker for a given candidate rather than applying a fixed per-user blend.

A.4 Cold-Start Threshold Analysis

Table 9 extends the cold-start analysis in Section 5.3 across all four interaction thresholds for CROSSATTENTION. Two complementary trends emerge.

First, the *standardised* warm/cold gap narrows monotonically from $d = 1.066$ at $\tau = 5$ to $d = 0.709$ at $\tau = 20$, indicating that cold- and warm-start scores become statistically less separable as users accumulate interaction history. The raw NDCG@10 gap (warm−cold) widens slightly over the same range ($0.1935 \rightarrow 0.2531$): both cold and warm scores rise with τ, but within-group variance evidently grows enough that the standardised effect size still shrinks. The two are not in tension – d normalises by pooled standard deviation, so a widening raw gap and a narrowing d can coexist when variance grows faster than the mean difference.

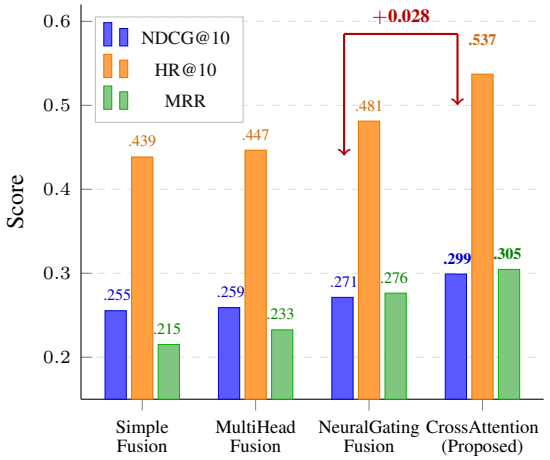


Figure 3: Fusion architecture ablation across NDCG@10, HR@10, and MRR. All variants share the same Dual-Encoder backbone and training data. Scores reflect the candidate-specific CrossAttention gate. All differences vs. CrossAttention are significant at $p<0.001$ (bootstrap n=1,000; Wilcoxon).

Second, and importantly, *cold* NDCG@10 itself rises monotonically with τ ($0.2834 \rightarrow 0.3444$, a $+21.5\%$ relative improvement), demonstrating that performance improves continuously as users engage further with the system even before reaching full warm-start status. This trajectory confirms that effective personalisation emerges well within the first few weeks of use, and that the initial cold-start penalty is transient rather than permanent.

A.5 Human Evaluation Results

Table 10 reports the full results of the two human evaluation studies described in §5.5.

A.6 Qualitative Examples

The representative examples below (Table 12) are drawn consistent with these aggregate statistics.

The full ten-example set spans all four behavioural outcomes of the CoVe pipeline: large revision, moderate revision, no revision, and negative revision providing a representative picture of verifier behaviour across diverse paper topics. Negative revision cases arise when the verifier flags trivially true statements (e.g., "Paper [1] is a research paper"), causing over-correction of an already partially faithful explanation, or when the baseline was conservative enough that any rewrit-

Table 8: Latency and cost breakdown for a single HYBRIDSCHOLAR user session (top-$K = 10$ recommendations). Latency measured on a single NVIDIA T4 GPU (Kaggle environment). Wall-clock assumes parallel explanation generation across all K candidates; sequential upper bound is $\approx$50 s. GPT-4o-Mini cost computed at current pricing ($0.15 / 1M input tokens, $0.60 / 1M output tokens); 6 verification calls per session, $\approx$600 input + 150 output tokens each.

Component	Latency	Notes
Ranking pipeline		
Dual-Encoder inference	419 ms	T4 GPU; scores all 239,618 papers
FAISS retrieval (top-5)	25 ms	GPU Flat L2, L2-normalised index
CrossAttention fusion	$<$1 ms	In-memory gate forward pass
Explanation pipeline (per candidate)		
Llama-3.1-8b generation	$\approx$2,000 ms	Groq API, $T = 0.3$
GPT-4o-Mini CoVe verify	$\approx$1,500 ms	OpenAI API, $T = 0.0$
GPT-4o-Mini CoVe revise	$\approx$1,500 ms	OpenAI API; only when claims flagged
Wall-clock per session[†]	$\approx$6,000 ms	Parallel generation across $K = 10$
Sequential (no parallelism)	$\approx$50,000 ms	Upper bound
Cost per 1,000 user sessions		
Dual-Encoder + FAISS (GPU)	$0.00	In-house GPU infrastructure
Llama-3.1-8b (Groq)	$1.00	$0.05 / 1M tokens
GPT-4o-Mini CoVe (6 calls)	$1.08	$0.00018 / call $\times$ 6 $\times$ 1,000
Total	$2.08	Per 1,000 sessions
FAISS index maintenance		
New arXiv CS papers/day	$\approx$200	Estimated from arXiv CS submission rate
Flat L2 rebuild time	$\approx$1.5 min	CPU; 239,618 vectors, dim = 256
ANN alternative (HNSW)	10$\times$ faster	$<0.1\%$ recall loss vs. Flat L2

[†] Wall-clock of $\approx$6 s assumes all $K = 10$ explanation generation calls are dispatched in parallel (asynchronous API requests). Distilling the GPT-4o-Mini verifier into a locally-deployable model (e.g. Llama-3.1-8b with CoVe fine-tuning) would eliminate the $1.08 API cost per 1,000 sessions entirely; this remains a direction for future work (§7).

Table 9: CrossAttention cold or warm NDCG@10 across four thresholds $\tau \in \{5, 10, 15, 20\}$, computed with the corrected candidate-specific gate and live s_{RAG} (§3.2.3). Gap = warm − cold. Both cold and warm NDCG@10 rise monotonically with τ. $^{***}p<0.001$ (Wilcoxon).

τ	Cold	Hit$_c$	Warm	Hit$_w$	Gap	d
5	0.2834	0.4492	0.4769	0.4968	+0.1935	1.066***
10	0.3238	0.5095	0.5279	0.5502	+0.2041	0.983***
15	0.3282	0.5220	0.5549	0.5742	+0.2267	0.882***
20	0.3444	0.5560	0.5975	0.6307	+0.2531	0.709***

ing increases the hallucination score; both failure modes reflect a known limitation of atomic decomposition verifiers that lack a notion of claim salience. No-revision cases confirm that faithfulness gains can occur through implicit generation conditioning without explicit flagging. Future work should explore salience-aware claim filtering and confidence thresholds to reduce over-correction while preserving the large- and moderate-revision gains.

Table 10: Human evaluation results. Study 1: expert annotation of 100 blind pairs (3 annotators; mean pairwise Cohen's κ). Study 2: end-user forced-choice preference (12 participants $\times$ 6 scenarios; n=72 decisive trials; h = Cohen's h). $^{***}p<0.0001$ (sign test, Study 2); consistent across all annotators (Study 1).

Study	Dimension	CoVe	RAG	Δ	Agreement	Sig.
Study 1 Expert (n=100)	Factual Acc. (%)	76.7	36.0	+40.7 pp	κ=0.72	***
	Usefulness (1–5)	4.39	3.99	+0.40	κ=0.48	***
	Naturalness (1–5)	4.47	4.24	+0.23	κ=0.39	***
	Pref. (CoVe/RAG/tie)	81	16	3 ties	—	—
Study 2 User (n=72)	Helpfulness Q1 (%)	78.8	21.2	+57.6 pp	h=0.614	***
	Trustworthiness Q2 (%)	79.3	20.7	+58.6 pp	h=0.626	***

Table 11: CoVe pipeline statistics across n=500 pairs. Verifier: GPT-4o-Mini. Hallucination rate = 1− faithfulness.

Metric	Baseline	+CoVe
Pipeline Activity		
Claims verified	11,371	
Claims flagged	1,117 (9.8%)	
Explanations revised	418/500 (83.6%)	
Hallucination Rate ↓		
MiniCheck (primary)	0.348	**0.265**
NLI (cross-val.)	0.802	**0.725**
Relative Reduction		
MiniCheck	−23.8% (***)	
NLI	−9.6% (***)	
Scorer Agreement		
Spearman ρ	0.138 (p<0.001)	

Table 12: CoVe qualitative examples across all behavioural categories, including the three representative cases from Section 5.6 (GPT-4o-Mini verifier). H_{base} and H_{CoVe} are MiniCheck hallucination scores (lower = more faithful). $\Delta = H_{\text{base}} - H_{\text{CoVe}}$ (positive = improved). Titles truncated for space. †No claims flagged; reduction driven by verification pass conditioning more conservative generation.

Paper	RAG	Action	H_{base}	H_{CoVe}	Δ	GPT-4o-Mini: flagged claim
Large Revision						
CNN-SLAM: Real-time Dense Monocular SLAM	0.70	Revised	0.639	0.056	+0.583	"The technique is used for real-time dense depth estimation"
Attention-Enhanced Lightweight Hourglass Network for Human Pose Estimation	0.70	Revised	0.964	0.437	+0.528	"The proposed method offers a novel approach to human pose estimation" (3 claims flagged)
Learning to Adapt SAM for Segmenting Cross-domain Point Clouds	0.72	Revised	0.574	0.047	+0.527	"The Self-Attention Mechanism (SAM) is used for segmenting data" (5 claims flagged)
Moderate Revision						
Denotational Foundations for Expected Cost Analysis	0.76	Revised	0.612	0.364	+0.248	"Cost can be analysed using various frameworks"
Enhancing Semantic Consistency of Large Language Models through Model Editing	0.77	Revised	0.419	0.171	+0.247	"Finetuning with prompt-output pairs improves consistency" (3 claims flagged)
Intriguing Properties of Large Language and Vision Models	0.77	Revised	0.300	0.055	+0.245	"MLLMs are capable of playing games" (5 claims flagged)
No Revision						
SpaRP: Fast 3D Object Reconstruction from Sparse Views	0.64	None†	0.976	0.293	+0.683	—
Redefining Data Pairing for Motion Retargeting	0.75	None†	0.411	0.084	+0.327	—
Negative Revision						
Efficient Exploration with Object-Centric World Model Learning	0.70	Revised	0.139	0.610	−0.471	"Paper [1] is a research paper." (4 claims flagged, over-correction)
3D-CT-GPT: Generating 3D Radiology Reports	0.73	None	0.201	0.653	−0.452	— (0 claims flagged, unexpected degradation)