🎓 Scholar Inbox: Personalized Paper Recommendations for Scientists

Markus Flicke   Glenn Angrabeit   Madhav Iyengar   Vitalii Protsenko

Illia Shakun   Jovan Cicvaric   Bora Kargi   Haoyu He   Lukas Schuler

Lewin Scholz   Kavyanjali Agnihotri   Yong Cao   Andreas Geiger

University of Tübingen, Tübingen AI Center

www.scholar-inbox.com

Abstract

Scholar Inbox is a new open-access platform designed to address the challenges researchers face in staying current with the rapidly expanding volume of scientific literature. We provide personalized recommendations, continuous updates from open-access archives (arXiv, bioRxiv, etc.), visual paper summaries, semantic search, and a range of tools to streamline research workflows and promote open research access. The platform's personalized recommendation system is trained on user ratings, ensuring that recommendations are tailored to individual researchers' interests. To further enhance the user experience, Scholar Inbox also offers a map of science that provides an overview of research across domains, enabling users to easily explore specific topics. We use this map to address the cold start problem common in recommender systems, as well as an active learning strategy that iteratively prompts users to rate a selection of papers, allowing the system to learn user preferences quickly. We evaluate the quality of our recommendation system on a novel dataset of 800k user ratings, which we make publicly available, as well as via an extensive user study.
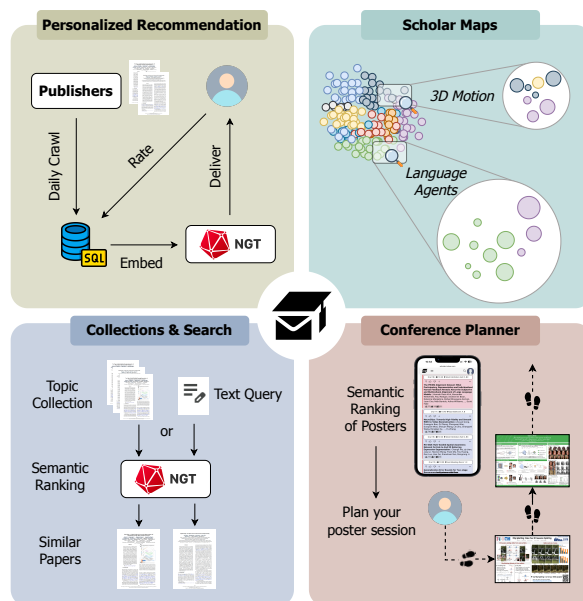
Figure 1: Key features of Scholar Inbox, including *Personalized Recommendations* tailored to individual interests, *Scholar Maps* for cross-domain paper exploration, *Collections* for literature review and exploration of new research areas, and *Conference Planner* for efficient time prioritization at conference poster sessions.

1   Introduction

The exponential growth of scientific publications has posed significant challenges for both junior and senior researchers to stay up-to-date with the latest relevant works (Fortunato et al., 2018; Zheng et al., 2024). This motivated the development of academic recommenders, which offer personalized paper recommendation services, aiming to promote the discovery of relevant works and enhance the efficiency of the research cycle.

However, despite these efforts, current platforms often fail to fully meet user requirements. For example, many researchers rely on platforms like X[1],

ResearchGate[2] or LinkedIn[3] for paper recommendations, which implicitly introduce biases towards popular authors and institutions via the Matthew effect (Perc, 2014; Färber et al., 2023). Furthermore, where personalized recommendations are offered, they are typically based on broadly defined topics (Wang, 2025), leading to an inaccurate understanding of user interests and thus suboptimal paper recommendations (Li et al., 2021).

In this paper, we present Scholar Inbox, a publicly available open-access platform with more accurate personalized recommendations and a wide range of functionalities for researchers, aiming

---

[1] www.x.com

[2] www.researchgate.net
[3] www.linkedin.com

to enhance research efficiency and promote open-access publications. As shown in Fig. 1, the advantages of Scholar Inbox primarily include four aspects: **(1) Personalized Recommendations:** We train a recommendation model for each researcher based on their positive and negative ratings during registration and while visiting our website. Unlike social media recommendations, our recommendations are only based on the paper content and therefore unbiased by social factors. **(2) Scholar Maps:** To facilitate exploration of papers across domains, we project all papers into a two-dimensional space based on their semantic representations, allowing users to easily search and discover research. **(3) Collections and Search:** We enable users to explore papers that are semantically similar to their collections and search similar papers based on free-form text descriptions. **(4) Conference Planner:** For large conferences, we offer a planner that helps users prioritize their time at poster sessions.

Besides offering a range of functionalities, we introduce a content-based recommendation model for research papers, provide a demonstration video[4], and release our dataset[5] of anonymized user ratings to support and facilitate future research on scientific recommender systems. In the following sections, we summarize existing academic platforms (§2), present the system architecture of Scholar Inbox (§3), and provide comprehensive evaluations, demonstrating its ability to deliver better recommendations and enhance user satisfaction (§4).

2 Related Work

Scientific Paper Recommendation Platforms: To meet growing research demands, support systems such as search engines, exploratory tools, and recommenders have emerged. Search engines like Google Scholar and Semantic Scholar rely on user-provided keywords. Research interests are however often multi-faceted and many new researchers are unaware of which terms accurately describe their desired search results. Exploratory tools such as Connected Papers[6] and Research Rabbit[7] fill this gap by visualizing citation graphs as 2D maps to show related papers to the user. Additionally, semantic paper maps of research have been created using t-SNE (González-Márquez et al., 2022).

---

[4]https://youtu.be/4fgM-iJgXJs
[5]github.com/avg-dev/scholar_inbox_datasets
[6]www.connectedpapers.com
[7]www.researchrabbit.ai

Recommendation Algorithms: Beyond exploration, researchers must read the latest research to stay relevant in their field and to avoid duplicate research. A plenitude of research recommenders have been proposed, but no system has so far achieved widespread adoption. Content-based filtering (CB) recommendation systems (Karpathy, 2025; Wang et al., 2018; Patra et al., 2020; Kart et al., 2022) generate recommendations purely using item information, but have been refined to include user interactions (Mohamed et al., 2022; Guan et al., 2010) and bibliographic information (Ma et al., 2021; Wang et al., 2018). Many implementations prefer sparse Term Frequency Inverse Document Frequency (TF-IDF) (Jones, 1972) embeddings over dense learning-based embeddings, due to their simplicity and lower runtime (Zhang et al., 2023; Hassan et al., 2019). Our ablation study corroborates that TF-IDF performs well for the research recommendation task, however we find that state-of-the-art distributed representations such as GTE (Li et al., 2023) outperform sparse embeddings in terms of vote prediction accuracy.

A known limitation of CB recommendation systems is the filter bubble effect (Portenoy et al., 2022) and diversity, novelty and serendipity have been identified as current limitations (Kreutz and Schenkel, 2022; Ali et al., 2021; Bai et al., 2019; Nguyen et al., 2014). In contrast, collaborative filtering (CF) derives recommendations from multiple users' interests and current approaches differ by whether they utilize author information (Utama et al., 2023; Neethukrishnan and Swaraj, 2017), use interactions (Murali et al., 2019; Xia et al., 2014) or bibliographic information (Sakib et al., 2020; Haruna et al., 2017; Liu et al., 2015).

Recent work focuses on hybrid systems, incorporating CB and CF into two-tower architectures (Church et al., 2024; Yi et al., 2019) or graph based approaches (Wang et al., 2024; Ostendorff et al., 2022; Cohan et al., 2020). CB, CF and hybrid approaches all suffer from the cold start problem for recommendation systems, as the recommender is uninformed about user preferences when they begin to use the system (Bai et al., 2019). There have been many attempts to alleviate this problem (Nura and Hamisu, 2024), for instance by uploading bibtex files from a reference manager (Kart et al., 2022). Scholar Inbox mitigates the cold start problem through a user-friendly onboarding process and an active learning strategy.

Research Recommendation Datasets: There are only a few research recommendation datasets available, such as Semantic Scholar Co-View (Cohan et al., 2020), SPRD (Sugiyama and Kan, 2010) and the largest dataset, CiteULike (Wang and Blei, 2011), contains 205k interactions. CiteULike's user-paper interaction are made when a user assigns a paper to their library, which implicitly shows that they liked that paper, but the exact reason why they added this paper is unclear. There is a lack of standard datasets in the field (Sharma et al., 2023), which is the reason we are releasing a dataset of 800k explicit positive/negative rating interactions from over 14.3k users. Furthermore, studies analyzing users' feedback to improve scholarly recommendation systems are rare and have very low number of responses (Zhang et al., 2023). We describe the outcomes of our user study with over 1.1k participants in the evaluation section.

3 Scholar Inbox

Our proposed scientific paper recommender system contains several key features, which we order by popularity according to our user survey:

Daily Digest: Daily paper updates (Fig. 4), ranked according to user interests provide a systematic way to keep up to date with research in the user's area of focus. The daily frequency of updates is designed to allow the user to build strong habits around staying informed in research.

Semantic Search: Users can search for papers by inserting free-form text. Example use-cases are to search for missed citations of related work sections, or to find papers that are similar to a paper the researcher is currently working on.

Conference Planner: Academic conferences are important for exchanging ideas, staying informed, and networking. To support this, we provide a poster session planner for leading machine learning conferences, which includes a personalized ranked list of posters and the ability to bookmark papers for later reading. We plan to extend this service to all scientific disciplines in the near future.

Collections: Any paper can be added to a user's collection for later reading. We show similar papers to each collection, such that the user can exploratively expand their collection.

Figure Previews: Along with the title, abstract and authors, we show the first five tables and figures of each paper, which we extract from the paper pdf using papermage (Lo et al., 2023).

3.1 Recommendation Model

To sort papers by relevance, Scholar Inbox uses a content-based recommender, which trains a logistic regression model on the user's paper ratings.

3.1.1 Training

Unlike traditional recommender systems that rely solely on implicit feedback from item interactions, Scholar Inbox enables users to tune their classifier through explicit up and downvotes. In addition to user ratings, we sample 5k random negative papers that the user has not interacted with, to better regularize the decision boundary. In contrast, our users have an average of 78 positive ratings, leading to a highly imbalanced dataset. To address this class imbalance, we use the weighted binary cross-entropy loss and assign distinct weights to positive ratings (w_P), negative ratings (w_N), and randomly sampled negatives (w_R):

$$\mathcal{L} = \frac{1}{n_T} \sum_{i=1}^{n_T} -w_i[y_i \log \hat{y}_i + (1 - y_i) \log (1 - \hat{y}_i)]$$

where n_T equals the total training set size. With n_P, n_N, and n_R representing the number of papers in each group, that is $n_T = n_P + n_N + n_R$, the weights of the two classes are balanced by:

$$n_P \, w_P = S \, (n_N \, w_N + n_R \, w_R) \qquad (1)$$

While the hyperparameter S controls the overall magnitude of negative weights (w_N and w_R), we introduce another hyperparameter V to adjust the relative importance between explicit negative ratings and randomly sampled negatives. For any chosen value of $V \in [0, 1]$, Eq. (1) is then satisfied using the following intermediate weights: $\tilde{w}_P = \frac{1}{n_P}$, $\tilde{w}_N = \frac{S \, V}{V \, n_N + (1-V) \, n_R}$, $\tilde{w}_R = \frac{S \, (1-V)}{V \, n_N + (1-V) \, n_R}$.

This formulation ensures that as users provide more explicit negative votes, the influence of randomly selected negatives on the overall weighting diminishes. However, it introduces a bias in the mean cross-entropy loss. Assuming each sample has an unweighted cross-entropy loss of 1, we derive:

$$\mathcal{L} = \frac{1}{n_T} \left(n_P \, \tilde{w}_P + n_N \, \tilde{w}_N + n_R \, \tilde{w}_R \right) = \frac{S + 1}{n_T}$$

This dependency on the total training set size n_T becomes problematic when applying weight decay and tuning the inverse regularization parameter C across users with different numbers of ratings. To correct for the bias, we multiply all final
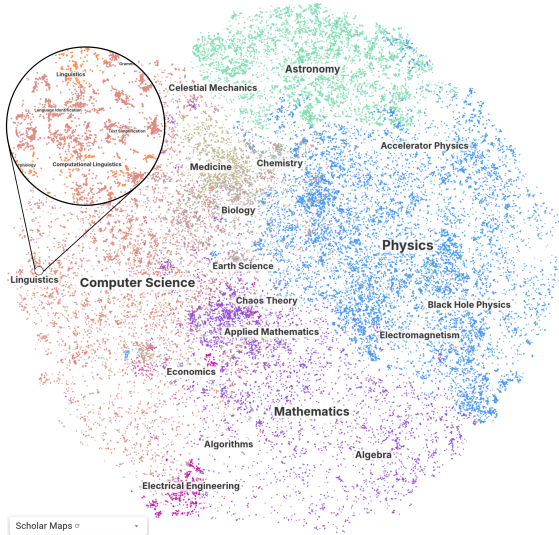
Figure 2: A t-SNE projection of the embedding space of all 3M papers in our database. The most cited papers and biggest topics are shown first. As the user zooms in, more papers are loaded dynamically.

Key Features	Google Scholar	Semantic Scholar	X	Emati	Arxiv Sanity	Research Rabbit	**Scholar Inbox**
Daily Recom.	✗	✗	✓	✗	✓	✓	✓
Multi-domain	✓	✓	✓	✗	✗	✓	✓
Non-redundant	✗	✓	✗	✓	✓	✓	✓
User ratings	✗	✗	✓	✓	✓	✗	✓
Lexical search	✓	✓	✓	✓	✓	✓	✓
Semantic search	✓	✗	✗	✗	✗	✗	✓
Collections	✓	✓	✗	✗	✓	✓	✓
Paper maps	✗	✗	✗	✗	✗	✓	✓
Dataset release	✗	✗	✗	✗	✗	✗	✓

Table 1: Comparison of features across research recommendation platforms, where *Daily Recom.* denotes daily recommendation, *User ratings* means the integration of user satisfaction metrics, and *Paper Maps* denotes the visualization of papers.

weights by n_T: $w_P = n_T \, \tilde{w}_P$, $w_N = n_T \, \tilde{w}_N$, and $w_R = n_T \, \tilde{w}_R$. Detailed ablation studies on the three hyperparameters C, V, and S are provided in the appendix. We linearly scale the output of our model to $[-100, 100]$ and display this relevance value for any paper on Scholar Inbox (Fig. 4).

3.1.2 Solving the Cold Start Problem

The cold start problem of recommender systems consists of the lack of user interaction history for new users. To provide an easy way to register to Scholar Inbox we offer users to add their own publications or publications from related authors via a simple author search. Alternatively, we allow users to navigate Scholar Maps, a 2D map of science, to quickly find relevant research fields and papers. We show a screenshot of `scholar-maps.com` in Fig. 2. The map is overlaid with topic labels, which we generated using Qwen (Qwen et al., 2025). We provide the prompt engineering strategies for label generation in the appendix. Labels are generated for four hierarchy levels (field, subfield, subsubfield, method), such that the field (Computer Science, Physics, etc.) is shown on the outermost zoom level. Subfields and method names of impactful papers are shown when zooming in, following Shneiderman's mantra "Overview first, zoom and filter, then details on demand" (Shneiderman, 1996). Once users find their research area, they select multiple papers that they are interested in. Users may search for papers by title or authors and

add papers that they like to their selection. In a second step, we provide an active learning framework, which employs stratified sampling, prioritizing papers near and above the recommender's decision boundary, and prompts the user to rate them. The recommender trains again after each rating is submitted, leading to iterative improvements.

3.2 User Centric Design

Most design decisions and features are first conceived by our users, before they are implemented by us. To reiterate the user focus, solicit user feedback and to make certain that Scholar Inbox addresses the concerns of its users, we regularly conduct user surveys.

As shown in Fig. 4, our website design follows a flat information hierarchy to minimise the number of clicks required to navigate to the desired functionality. The regular nature of our digest updates provides a habit forming experience, allowing our users to integrate Scholar Inbox into their daily work routine. We show a comparison of our features with other websites that recommend papers to researchers in Tab. 1.

3.3 Software Architecture

Fig. 3 shows the data processing pipeline. Scholar Inbox downloads papers and their metadata from preprint servers such as arXiv, bioRxiv, chemRxiv and medRxiv as well as directly from public conference proceedings. We compare and update missing fields in our database using the Semantic Scholar Open Research corpus (S2ORC) (Lo et al., 2020), to ensure that all papers are assigned the correct conference or journal upon publication. We also incorporate author information and the citation graph
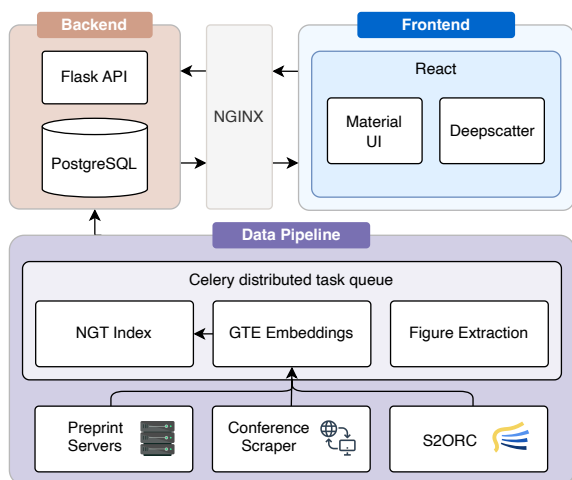
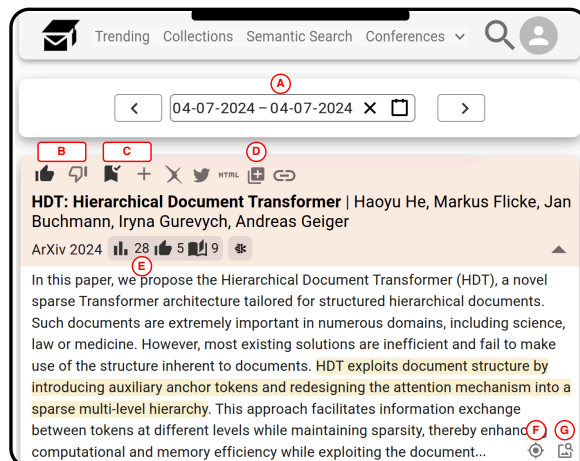Figure 3: Data flow through our processing pipeline.

Figure 4: Tablet or mobile phone view of the daily digest. To enable faster skim-reading, we highlight the sentence that is most related to the idea of the research paper. In red circles, we show the (A) date picker, (B) thumbs up/down buttons, (C) bookmarking/collections buttons, (D) bibtex button, (E) paper relevance score, (F) similar papers button and (G) teaser figure button.

from S2ORC. We concatenate titles and abstracts, separated by a special [SEP] token, to encode each paper with GTE_{large} (Li et al., 2023), an efficient state-of-the-art transformer encoder trained with multi-stage contrastive learning. The paper embeddings are stored in NGT[8], a high performance nearest neighbor search index. We use Celery[9] to handle asynchronous tasks, including extracting figures and text embeddings. NGINX is used to serve the frontend static files and to proxy requests to the backend and our user interface is built with React[10]. Scholar Maps uses deepscatter[11] with tiled loading and GPU acceleration using WebGL to provide a smooth user experience.

3.4 Daily Digest

The daily digest, as shown in Fig. 4, is the main feature of Scholar Inbox. It holds a ranked list of papers within a short time period (day or week) with title, abstract, authors and publication venue for each paper. Digest papers are ordered by their predicted relevance for the current user, which also determines the paper header's background color. Users may refine their recommendation model by rating papers positively or negatively using the thumbs buttons (B). Using a button, each paper shows images of figures and tables, as well as the option to show a list of semantically similar papers. Moreover, users can search for semantically similar papers (F) and preview a paper's figures and tables (G) with a single click. Papers can also be

---
[8] www.github.com/yahoojapan/NGT
[9] https://docs.celeryq.dev
[10] www.reactjs.org
[11] www.github.com/nomic-ai/deepscatter

bookmarked or added to collections (C), posted on social media or exported as bibtex to reference managers (D). In addition to viewing daily digests, the user may also aggregate relevant papers over a longer time range (A) and specify the weekdays on which to receive their digests via email. If a user returns to the site after an extended period of time, we provide a catch-up digest containing the most relevant papers during their time of absence.

4 Evaluation

4.1 Recommendation model

Model	Dim.	F1	nDCG	Balanced acc.	AUC
TF-IDF	10k	83.60 ±0.10	**88.67** ±0.29	75.74 ±0.05	84.41 ±0.09
TF-IDF	256	81.03 ±0.17	83.37 ±0.26	74.52 ±0.10	82.28 ±0.04
SPEC2	256	83.22 ±0.16	84.21 ±0.31	78.16 ±0.07	86.36 ±0.09
GTE-B	256	84.16 ±0.11	85.42 ±0.28	77.92 ±0.08	86.24 ±0.05
GTE-L	256	**84.51** ±0.15	85.83 ±0.22	**78.31** ±0.12	**86.75** ±0.07

Table 2: Performance of the recommender using different embedding methods. TF-IDF 10k is sparse with 10K dimensions, while the other models are dense and compressed to 256 dimensions using PCA.

We evaluate classic sparse (TF-IDF) and neural network-based dense (GTE, SPECTER2) embedding models for encoding research papers, measuring performance through two distinct approaches in Tab. 2. First, we follow established methodologies for recommender systems without explicit negative ratings (He et al., 2017) and evaluate each positive
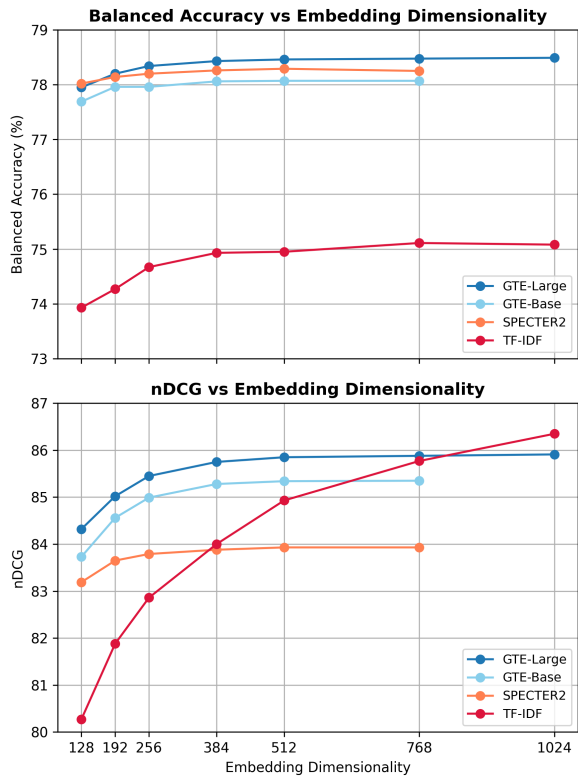
Figure 5: Performance of different embeddings after dimensionality reduction from their original sizes: GTE(1024), SPECTER(768), TF-IDF(10k). At its orginial dimensionality of 10k, TF-IDF achieves a score of 88.2 on nDCG.
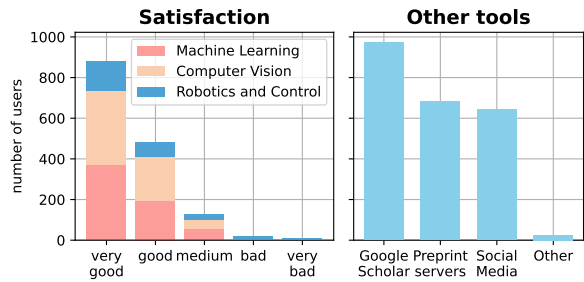

Figure 6: User study for Scholar Inbox among 1,233 participants. Left: Satisfaction levels across users in Machine Learning, Computer Vision, and Robotics indicate a highly positive experience. Right: Users also use search engines, preprint servers, and social media, but few rely on other recommender systems, underscoring Scholar Inbox's central role in paper discovery.

sample together with randomly sampled negative examples. For these, we compute F1-score and nDCG using a leave-one-out strategy for positively voted validation papers. While this evaluation is widely adopted in the literature, it fails to account for the impact of hard negative samples. We further analyze model performance including explicit negative user ratings on binary classification metrics (balanced accuracy and AUC) and find that GTE outperforms TF-IDF on classification between positives and hard negatives.

Evaluating qualitatively, we find GTE underperforms on nDCG primarily for two reasons: It assigns higher probabilities to sampled negatives that resemble users' positive training examples, and it assigns lower probabilities to certain positive validation papers which are also classified negatively by TF-IDF. The first case is susceptible to noise and the second has minimal impact on the digest, as neither model recommends these false negatives. Therefore, we select the GTE-Large model for its superior performance on explicit user ratings, which we consider

more reliable. Empirically, we also find that our dense embeddings yield better calibrated cosine similarities which benefit similar paper/semantic search and 2D visualizations like Scholar Maps.

Step further, we investigate the effect of PCA-based dimensionality reduction on transformer-derived embeddings for the recommendation task, as shown in Fig. 5. Performance is evaluated in terms of balanced accuracy (top) and nDCG (bottom) across varying embedding dimensions. For all transformer-based methods (GTE-Large, GTE-Base, and SPECTER2), both metrics remain relatively stable when reducing dimensions from 1024 to 256, suggesting redundancy in the original representations. Below 256, however, a notable degradation in performance emerges, indicating that further compression removes informative components. We conclude that not all dimensions are used efficiently for our recommendation task. Notably, TF-IDF exhibits steady gains with increased dimensionality. For runtime and memory efficiency we choose a dimensionality of 256 for the final GTE-large model.

4.2 User Study

To evaluate Scholar Inbox, we conduct a user study with 1233 participants, who are asked to rate their satisfaction with the platform on a scale from 1 to 5 in terms of usability, satisfaction, and the quality of recommendations. Their evaluation of Scholar Inbox is extremely positive, as can be seen from user voting in Fig. 6 and from the user retention statistics in Fig. 7(a). The most common criticism from our user study is that the platform currently does

312

(a) User Retention   (b) User Domain Distribution   (c) Paper Category Distribution
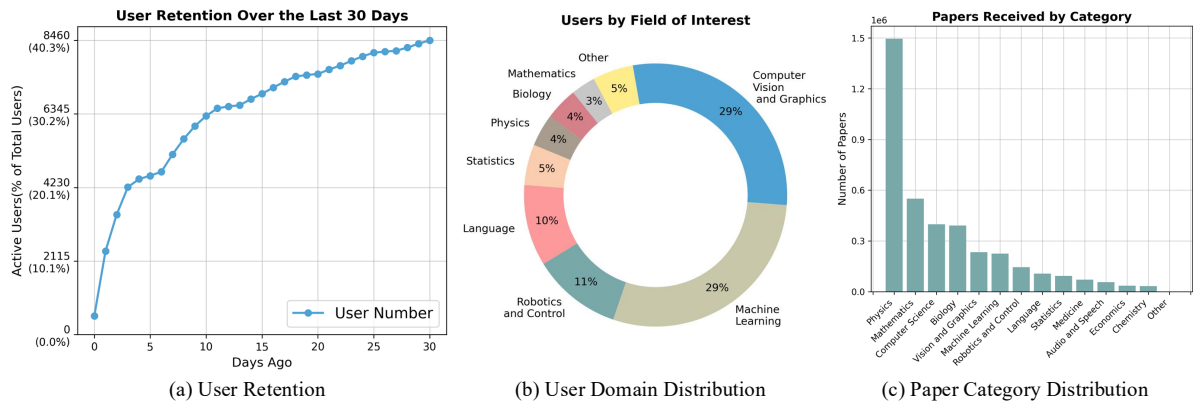
Figure 7: Statistics of user and papers in Scholar Inbox: (a) Cumulative number of active users in the past 30 days, demonstrating a high retention rate; (b) Distribution of users by research domain, indicating strong representation in ML and CV while maintaining multidisciplinary reach; and (c) Distribution of recommended papers by category, reflecting user interests across diverse scientific fields.

not support explicit modeling of separate research interests. Whilst we observe that multiple research interests are already handled well in a single recommender, we are working on enabling users to explicitly switch between different research interests in the next version of Scholar Inbox.

4.3 User Retention and Distribution

In Fig. 7(a), we present the cumulative number of users active in the last 30 days. This graph only shows user interactions on the website, excluding users who only read our email newsletter. Even though the number of registered users on Scholar Inbox is only 23k, which is relatively few for a website, 8k (35%) of them were active in the last 30 days. This high retention rate underscores both the effectiveness of the recommendation system and the practical value offered by the platform.

As shown in Fig. 7(b), while a significant portion of users focus on Machine Learning and Computer Vision, the presence of users from diverse fields such as Physics, Biology, Language, and Statistics demonstrates that our platform is attracting a broad range of researchers. This suggests its potential to effectively support interdisciplinary research across multiple scientific domains.

5 Conclusion

Scholar Inbox is a new open-access platform that provides daily, personalized recommendations for research papers and a range of tools to improve research workflows and promote open access to research. Our evaluation on a dataset of 800k user ratings and the user study highlight the platform's

effectiveness in providing accurate recommendations and enhancing user satisfaction.

Ethical Consideration

The data source used in Scholar Inbox is primarily constructed from publicly available metadata, i.e., arXiv and Semantic Scholar. Both sources explicitly permit data usage for research and non-commercial purposes under their respective terms of service. We ensured full compliance with these guidelines during the data acquisition process. Besides, our published dataset does not include any sensitive or user-specific information, thereby minimizing potential privacy risks and ethical concerns.

Our objective is to facilitate scholarly discovery and recommendation across a broad spectrum of scientific disciplines. To this end, we curated a domain-diverse dataset that reflects the interdisciplinary nature of contemporary research. As illustrated in Fig. 7(c), our papers spans a wide range of fields up to the time of this publication. This disciplinary breadth reflects our vision to support researchers across diverse academic fields.

References

Zafar Ali, Irfan Ullah, Amin Khan, Asim Ullah Jan, and Khan Muhammad. 2021. An overview and evaluation of citation recommendation models. *Scientometrics*, 126(5):4083–4119.

Xiaomei Bai, Mengyang Wang, Ivan Lee, Zhuo Yang, Xiangjie Kong, and Feng Xia. 2019. Scientific Paper Recommendation: A Survey. *IEEE Access*, 7:9324–9339. Conference Name: IEEE Access.

Kenneth Church, Omar Alonso, Peter Vickers, Jiameng Sun, Abteen Ebrahimi, and Raman Chandrasekar. 2024. Academic Article Recommendation Using Multiple Perspectives. *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. Association for Computational Linguistics.

Michael Färber, Melissa Coutinho, and Shuzhou Yuan. 2023. Biases in scholarly recommender systems: impact, prevalence, and mitigation. *Scientometrics*, 128(5):2703–2736.

Santo Fortunato, Carl T Bergstrom, Katy Börner, James A Evans, Dirk Helbing, Staša Milojević, Alexander M Petersen, Filippo Radicchi, Roberta Sinatra, Brian Uzzi, et al. 2018. Science of science. *Science*, 359(6379):eaao0185.

Rita González-Márquez, Philipp Berens, and Dmitry Kobak. 2022. Two-dimensional visualization of large document libraries using t-SNE. In *Proceedings of Topological, Algebraic, and Geometric Learning Workshops 2022*, volume 196 of *Proceedings of Machine Learning Research*, pages 133–141. PMLR.

Ziyu Guan, Can Wang, Jiajun Bu, Chun Chen, Kun Yang, Deng Cai, and Xiaofei He. 2010. Document recommendation in social tagging services. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 391–400, New York, NY, USA. Association for Computing Machinery.

Khalid Haruna, Maizatul Akmar Ismail, Damiasih Damiasih, Joko Sutopo, and Tutut Herawan. 2017. A collaborative approach for research paper recommender system. *PLOS ONE*, 12(10):e0184516. Publisher: Public Library of Science.

Hebatallah A. Mohamed Hassan, Giuseppe Sansonetti, Fabio Gasparetti, Alessandro Micarelli, and J. Beel. 2019. Bert, elmo, use and infersent sentence encoders: The panacea for research-paper recommendation? In *Proceedings of ACM RecSys 2019 Late-breaking Results co-located with the 13th ACM Conference on Recommender Systems (RecSys 2019)*, volume 2431, pages 6–10.

Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, pages 173–182, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Karen Spärck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21.

Andrej Karpathy. Arxiv sanity preserver [online]. 2025.

Özge Kart, Alexandre Mestiashvili, Kurt Lachmann, Richard Kwasnicki, and Michael Schroeder. 2022. Emati: a recommender system for biomedical literature based on supervised learning. *Database*, 2022:baac104.

Christin Katharina Kreutz and Ralf Schenkel. 2022. Scientific paper recommendation systems: a literature review of recent publications. *International Journal on Digital Libraries*, 23(4):335–369.

Yi Li, Ronghui Wang, Guofang Nan, Dahui Li, and Minqiang Li. 2021. A personalized paper recommendation method considering diverse user preferences. *Decision Support Systems*, 146:113546.

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

Haifeng Liu, Xiangjie Kong, Xiaomei Bai, Wei Wang, Teshome Megersa Bekele, and Feng Xia. 2015. Context-Based Collaborative Filtering for Citation Recommendation. *IEEE Access*, 3:1695–1703. Conference Name: IEEE Access.

Kyle Lo, Zejiang Shen, Benjamin Newman, Joseph Chang, Russell Authur, Erin Bransom, Stefan Candra, Yoganand Chandrasekhar, Regan Huff, Bailey Kuehl, Amanpreet Singh, Chris Wilhelm, Angele Zamarron, Marti A. Hearst, Daniel Weld, Doug Downey, and Luca Soldaini. 2023. PaperMage: A unified toolkit for processing, representing, and manipulating visually-rich scientific documents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 495–507, Singapore. Association for Computational Linguistics.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online. Association for Computational Linguistics.

Shutian Ma, Heng Zhang, Chengzhi Zhang, and Xiaozhong Liu. 2021. Chronological citation recommendation with time preference. *Scientometrics*, 126(4):2991–3010.

Hebatallah A. Mohamed, Giuseppe Sansonetti, and Alessandro Micarelli. 2022. Tag-Aware Document Representation for Research Paper Recommendation. *arXiv preprint*. ArXiv:2209.03660.

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. ISSN: 2575-7288.

K. V. Neethukrishnan and K. P. Swaraj. 2017. Ontology based research paper recommendation using personal ontology similarity method. In *2017 Second International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, pages 1–4.

Tien T. Nguyen, Pik-Mai Hui, F. Maxwell Harper, Loren Terveen, and Joseph A. Konstan. 2014. Exploring the filter bubble: the effect of using recommender systems on content diversity. In *Proceedings of the 23rd international conference on World wide web*, WWW '14, pages 677–686, New York, NY, USA. Association for Computing Machinery.

Mukhtar Nura and Zaharaddeen Adamu Hamisu. 2024. An author-centric scientific paper recommender system to improve content-based filtering approach. *International Journal of Software Engineering and Computer Systems*, 10(1):40–49. Number: 1.

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. *Journal of Biomedical Informatics*, 104:103399.

Matjaž Perc. 2014. The matthew effect in empirical data. *Journal of The Royal Society Interface*, 11(98):20140378.

Jason Portenoy, Marissa Radensky, Jevin D West, Eric Horvitz, Daniel S Weld, and Tom Hope. 2022. Bursting Scientific Filter Bubbles: Boosting Innovation via Novel Author Discovery. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, pages 1–13, New York, NY, USA. Association for Computing Machinery.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 Technical Report. *arXiv preprint*. ArXiv:2412.15115 [cs].

Nazmus Sakib, Rodina Binti Ahmad, and Khalid Haruna. 2020. A Collaborative Approach Toward Scientific Paper Recommendation Using Citation Context. *IEEE Access*, 8:51246–51255. Conference Name: IEEE Access.

Ritu Sharma, Dinesh Gopalani, and Yogesh Meena. 2023. An anatomization of research paper recommender system: Overview, approaches and challenges. *Engineering Applications of Artificial Intelligence*, 118:105641.

Ben Shneiderman. 1996. The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations.

Kazunari Sugiyama and Min-Yen Kan. 2010. Scholarly paper recommendation via user's recent research interests. In *Proceedings of the 10th annual joint conference on Digital libraries*, pages 29–38, Gold Coast Queensland Australia. ACM.

Ferzha Putra Utama, Triska Mardiansyah, Ruvita Faurina, and Arie Vatresia. 2023. Scientific articles recommendation system based on user's relatedness using item-based collaborative filtering method. *Jurnal Teknik Informatika (Jutif)*, 4(3):467–475. Number: 3.

Chang Wang. Paper digest [online]. 2025.

Chong Wang and David M. Blei. 2011. Collaborative topic modeling for recommending scientific articles. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 448–456, San Diego California USA. ACM.

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

Le Wang, Wenna Du, and Zehua Chen. 2024. Multi-Feature-Enhanced Academic Paper Recommendation Model with Knowledge Graph. *Applied Sciences*, 14(12):5022. Number: 12 Publisher: Multidisciplinary Digital Publishing Institute.

Feng Xia, Nana Yaw Asabere, Haifeng Liu, Nakema Deonauth, and Fengqi Li. 2014. Folksonomy based socially-aware recommendation of scholarly papers for conference participants. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14 Companion, pages 781–786, New York, NY, USA. Association for Computing Machinery.

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*, RecSys '19, pages 269–277, New York, NY, USA. Association for Computing Machinery.

Zitong Zhang, Braja Gopal Patra, Ashraf Yaseen, Jie Zhu, Rachit Sabharwal, Kirk Roberts, Tru Cao, and Hulin Wu. 2023. Scholarly recommendation systems: a literature survey. *Knowledge and Information Systems*, 65(11):4433–4478.

Yuxiang Zheng, Shichao Sun, Lin Qiu, Dongyu Ru, Cheng Jiayang, Xuefeng Li, Jifan Lin, Binjie Wang, Yun Luo, Renjie Pan, Yang Xu, Qingkai Min, Zizhao Zhang, Yiwen Wang, Wenjie Li, and Pengfei Liu. 2024. OpenResearcher: Unleashing AI for accelerated scientific research. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 209–218, Miami, Florida, USA. Association for Computational Linguistics.

A   Appendix

A.1   Prompt Engineering Strategies for t-SNE Label Generation

To extract the topic hierarchy for t-SNE visualization, we conducted LLM inference on each paper using a prompt composed of four distinct parts: Task, Additional Note, Format, and Title & Abstract. The Task section provides the general extraction instructions and mandates strict adherence to the specified format while explicitly instructing the model to omit any additional commentary to simplify output parsing. The Additional Note section restricts the field values to a predefined, handcrafted list of scientific disciplines. The Format section details the precise structure of the expected output along with explanations of the corresponding fields. Finally, the Title & Abstract section contains the actual text to be processed for extracting the required information.

During prompt engineering, we determined that including the format explanation only once, positioned as late as possible before the data, is optimal. Moreover, employing an explicit empty field placeholder proved crucial for smaller LLMs, as it enhances structural consistency and prevents unnecessary repetitions in the output.

```
Task: Based on the title and abstract provided, extract
and label the following key details exactly as specified:
field_of_Paper, subfield, sub_subfield, keywords, method_
name_shortname. Follow the structure exactly and keep your
answers brief and specific. Adhere strictly to the format.
If any information is unclear or unavailable in the abstract,
write "None." for that field. Use the exact labels and
```

```
formatting provided. Do not include comments or repeat any
part of the response. Note: For field_of_Paper, choose one
from the following list of academic disciplines:
Mathematics, Physics, Chemistry, ...

Details to Extract:
field_of_Paper =
*The primary academic discipline from the list above.*
[insert answer]
subfield =
*The main research category within the field.*
[insert answer]
sub_subfield =
*A narrower focus within the subfield.*
[insert answer]
keywords =
*A set of 3-5 words or phrases that describe the core topics,
separated by commas.*
[insert answer]
method_name_shortname =
*The main technique or model name proposed in the abstract.*
[insert answer]

Title: [title]
Abstract: [abstract]
```

Listing 1: Scholar Map's label generation prompt. For better readability, we shortened the list of disciplines.

A.2   Technical Challenges

Extracting teaser figures (or getting GTE embeddings) is compute-intensive; however, leveraging GPU acceleration facilitates rapid inference and efficient parallel processing of papers. For efficiency our architecture enables external machines to connect to the main server's broker and backend (powered by Redis) via SSH port forwarding. This setup allows remote Celery workers to access tasks directly from the Scholar server. Consequently, any machine with the appropriate credentials—regardless of its physical location—can serve as a task consumer within our distributed environment, making our pipeline scalable by allowing us to seamlessly connect additional machines to accelerate computations as needed.

A.3   Hyperparameter Ablation Studies

We evaluate the sensitivity of our system to each of the three hyperparameters introduced in Section 3.1.1. For our ablation experiments, we use 256-dimensional GTE-Large embeddings with a standard configuration of $(C = 0.1, V = 0.8, S = 5.0)$. As in our main evaluation, balanced accuracy is calculated using explicit negative votes, while F1 and nDCG refer to 100 randomly sampled negatives. The results are summarized in Figure 8.

A.3.1   Inverse Regularization Strength C

With V and S fixed at their standard configuration values, positive weights w_P are higher than negative weights w_N and w_R. The model prioritizes fitting positive training examples, achieving the highest recall at $C = 10^{-1.5}$ (where F1 and
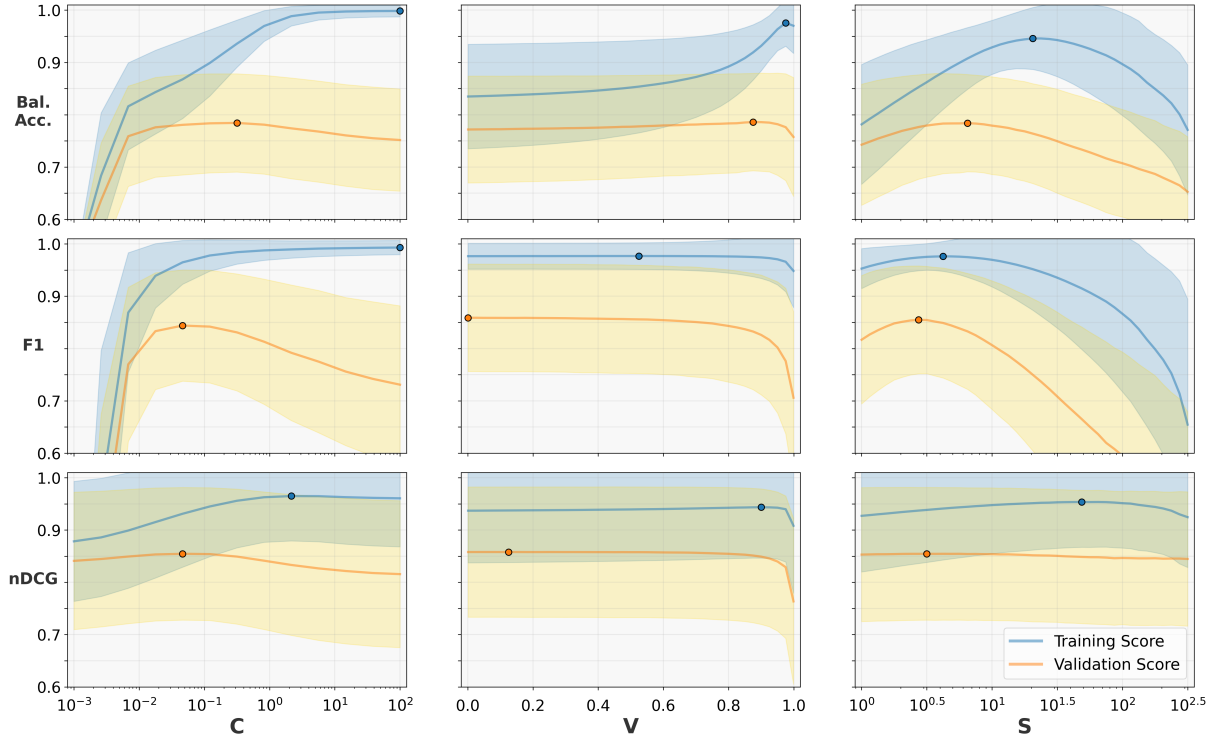
Figure 8: Hyperparameter ablation studies on GTE-Large embeddings. The metrics correspond to those in Table 2. Each plot shows the effect of individually varying one parameter while keeping the others fixed. Shaded regions indicate ± 1 standard deviation across the user base (not across random seeds).

nDCG are maximized). Further increasing C allows the model to better fit explicit negative examples, improving specificity and balanced accuracy (optimal at $C = 10^{-0.5}$). However, this tightens the decision boundary around difficult negatives, reducing performance between positives and simpler sampled negatives, consequently lowering F1 and nDCG. We note that linear classification applied to higher-dimensional embeddings contains a larger number of parameters and therefore attains similar performance under stronger regularization (e.g. $C = 0.05$ for 1024-dimensional GTE-Large).

A.3.2 Explicit-to-Random Negative Ratio V

The hyperparameter V controls the trade-off between performance on explicit negatives and randomly sampled negatives. Raising it from 0 to 0.9 elevates specificity on explicit negatives from 68% to 78% and maximizes balanced accuracy at 78.6% (up from 77.2%). The increased emphasis on difficult negative examples again tightens the decision boundary, producing false negatives and causing a monotonic decrease in F1 and nDCG. Nonetheless, we select a larger value $V = 0.8$ as it makes the model more receptive to downvotes and allows users to tune their classifier by explicitly stating

which papers should not be recommended to them.

A.3.3 Negative Weights Scale S

The hyperparameter S controls the magnitude of the negative weights w_N and w_R. At low values ($S = 1$), the model exhibits highly imbalanced class behavior with a recall of 94% but a specificity on explicit negatives of only 55%. Raising S mitigates this disparity, with all three metrics reaching high scores at our standard configuration value. As S increases, the model assigns progressively lower logits to all samples. Beyond $S = 5$, this reduction becomes substantial enough to cause a notable drop in recall, lowering balanced accuracy and F1. In contrast, nDCG remains stable up to much higher values ($S = 10^3$) because the model preserves the relative ranking between positives and randomly sampled negatives until positive weights become negligibly small compared to negative weights.