ScientiaRec: a Scientific Article Recommendation System with LLM-Driven Feature Extraction

Imen Ben Sassi
imen.ben-sassi@lirmm.fr
LIRMM, Université de Montpellier, CNRS
Montpellier, France

ABSTRACT

According to the European Commission, the number of scientific publications produced annually worldwide has more than tripled between 2000 and 2022. This rapid growth has made it increasingly difficult for researchers to identify and keep up with relevant work. Recommender systems (RSs) have emerged as a promising solution to this problem by helping users navigate the expanding scientific literature. In this paper, we introduce ScientiaRec, a novel scientific article RS that combines user-item interactions and content-based features, including descriptive tags and keywords automatically extracted using a large language model (LLM). At the core of our approach is a serendipity-aware matrix factorization model, designed to recommend relevant items while actively promoting serendipity. The goal is to help users discover novel and potentially insightful papers that go beyond their immediate research interests. We evaluate the performance of ScientiaRec against several baseline models using the publicly available CiteULike dataset, employing a comprehensive set of both accuracy-oriented and beyond-accuracy evaluation metrics. In addition, we conduct an LLM-based study to assess the serendipity of ScientiaRec Top-N recommendations. The experimental results demonstrate that ScientiaRec achieves a strong balance between relevance and beyond-accuracy objectives. Moreover, the inclusion of LLM-derived keywords significantly enhances the RS's overall performance.

CCS CONCEPTS

• **Information systems → Recommender systems**.

KEYWORDS

Recommender systems, Scientific article, Feature extraction, Large Language Model, Serendipity

ACM Reference Format:
Imen Ben Sassi. 2025. ScientiaRec: a Scientific Article Recommendation System with LLM-Driven Feature Extraction. In *Proceedings of Workshop on Evaluating and Applying Recommendation Systems with Large Language Models at RecSys '25 (EARL '2025).* ACM, New York, NY, USA, 11 pages.

1 INTRODUCTION

The goal of a scientific article recommender system (RS) is to assist researchers by providing personalized suggestions of relevant scholarly papers from a large corpus. Each user has a history of interactions with articles—such as downloads, reads, or citations—and the system must predict and rank new articles that the user is likely to find valuable. In this paper we propose a RS named ScientiaRec that aims to capture both collaborative signals from user behavior and content-based signals inherent to scientific articles, such as keywords, abstracts, and domain-specific metadata. The challenge lies not only in recommending relevant items but also in promoting serendipity to help users discover novel and potentially insightful papers beyond their immediate research interests. This helps to break the "filter bubble", where users are stuck in a limited circle of familiar topics [20]. Serendipity happens when the RS suggests something the user did not expect, from a different domain, but upon reading it, he realize: "Oh! This actually fits my work in a cool, unexpected way".

The authors in [26] extend Matrix Factorization (MF) with a novelty-aware regularization. Building on this idea, we propose a feature-aware Serendipity MF model that leverages both collaborative signals and semantic features. Our main contributions are:

- **Feature-aware Matrix Factorization:** We enhance classic MF by concatenating each item's latent vector with averaged embeddings from its tags and LLaMA3-extracted keywords.
- **LLM-based Semantic Enrichment:** We extract keywords from article titles and abstracts using LLaMA3, adding rich, fine-grained descriptors to item profiles.
- **Serendipity-aware Loss Function:** A Jaccard-based regularization term penalizes recommendations that are either too similar or too dissimilar to a user's history, encouraging suggestions that are surprising yet still relevant.
- **LLM-based Study:** We use GPT-4o to evaluate ScientiaRec Top-N recommendation in terms of serendipity.

The rest of the paper is organized as follows. We start in Section 2 by an overview of scientific papers, serendipity-based and LLM-enriched RSs. Next, we describe our serendipity-driven RS in Section 3. Section 4 details our experimental study, including the offline and LLM-based evaluations. Finally, Section 5 summarizes our paper and discuss our plans for future directions.

2 RELATED WORK

Recommender systems for scientific articles have gained significant attention due to the rapid growth of academic publications [8, 35]. These systems aim to assist researchers in discovering relevant literature. Existing works in scientific RSs can be broadly categorized into content-based filtering and collaborative filtering. We also give a quick review of serendipity-based and LLMs enhanced RSs.

EARL '2025, September 22–26, 2025, Prague, Czech Republic

2.1 Content-based Approaches

Content-based methods are the most commonly applied and are based on analyzing article metadata such as titles, abstracts, keywords, and text to recommend papers similar to those a user has previously read. The first recommendation system for research publications was introduced in the CiteSeer project and used content-based similarity methods [2]. The majority of the reviewed studies employed the Term Frequency-Inverse Document Frequency (TF-IDF) approach to assess text similarity. By down-weighting common terms, TF-IDF emphasizes more informative words within a document. To enhance both relevance and serendipity, Sugiyama et al. [24] generated feature vectors based on TF-IDF and built user profiles from the Co-Author Network. They then used cosine similarity to identify and recommend papers with higher similarity scores.

2.2 Collaborative Filtering Approaches

Collaborative filtering techniques leverage user-item interaction data, such as reading history or citation patterns, to make recommendations, and are generally divided into two main categories: model-based and memory-based methods. Model-based approaches rely on matrix factorization techniques, where user preferences are inferred through learned latent factor embeddings. In contrast, memory-based methods estimate user preferences by directly computing similarities—such as correlation coefficients or cosine similarity—between users or items using historical rating data. For example, to tackle the challenge of recommending newly published papers lacking citation history, the authors in [10] proposed a strategy based on Singular Value Decomposition (SVD) for matrix factorization and rating prediction. This approach allows the system to infer researchers' interests and recommend relevant papers even in the absence of citation data.

2.3 Serendipity-based Recommender Systems

Serendipity has emerged as a compelling concept within the field of RSs, drawing increasing attention from researchers [3, 34]. In this context, serendipity refers to a system's capacity to suggest items that are not only relevant but also pleasantly surprising—items that users are unlikely to discover on their own [5, 18]. A serendipitous recommendation is typically characterized by its novelty, unexpectedness, and relevance [14]. Relevance is generally measured by how closely an item aligns with the user's profile [6], while novelty implies that the item is unknown to the user and not easily found through conventional means [27]. However, assessing the degree of unexpectedness remains a challenging task. One attempt to incorporate serendipity into RSs was the content-based approach introduced in [11]. The authors used a supervised learning model based on item textual features to estimate the probability that a previously unseen item would be relevant to a specific user. In this framework, items about which the system is uncertain—neither clearly relevant nor irrelevant—are treated as potentially serendipitous and prioritized for recommendation. A different perspective was proposed in [1], where unexpected items are identified by first modeling what a user is likely to expect. Expected items include those already rated by the user and those similar to them; items falling outside this set are then considered unexpected. Additionally, the work in [17] introduced a RS for academic papers that leverages BisoNets to suggest items across two distinct domains, aiming to deliver serendipitous discoveries through domain-spanning recommendations. Recently, the authors in [23] explored the impact of serendipity on the quality of point-of-interest (POI) recommendations by introducing a novel POI RS called DISCOVERY, which aims to enhance the balance between accuracy and serendipity.

2.4 LLMs Enhanced Recommender Systems

Recent advancements in generative models have played a pivotal role in shaping the development of RSs [7]. The emergence of pretrained LLMs with powerful, generalized natural language reasoning capabilities has marked a new era in the recommendation domain. Over the past few years, a wave of models has shaped the integration of LLMs into RSs. A pioneering effort was BERT4Rec [25], which laid the groundwork for transformer-based sequential recommendation by modeling user behavior as masked sequences. Later models such as GPTRec [21] explored autoregressive paradigms for recommendation, leveraging the generative capabilities of GPT-2 architecture for next-item recommendations. As LLMs matured, models like LLaMARec [32] began to exploit the power of open-source LLMs such as LLaMA for domain-specific fine-tuning on recommendation tasks, demonstrating strong performance even with limited supervision via prompt engineering. ChatRec [9] marked a turning point by extending conversational recommender systems through dialogue-aware LLMs, enabling dynamic preference elicitation and interactive recommendation via natural language exchanges. A novel framework for narrative-driven recommendation, where users express their preferences through free-form natural language narratives, was proposed in [19]. Recognizing the scarcity of labeled narrative-query data, the authors propose leveraging a pretrained LLM (InstructGPT) to generate synthetic narrative queries from user interaction histories. These generated narratives, conditioned on user reviews and past items, are used to train a lightweight retrieval model. The authors in [12] addressed the problem of users with sparse activity or "weak" interaction histories by proposing a hybrid task allocation framework that couples traditional recommendation models with in-context learning using LLMs. The system first identifies weak users with below-threshold ranking performance and minimal engagement, then delegates their ranking tasks to an LLM, informed by their interaction history, while relying on standard RS outputs for the rest. A new prompting-based representation learning method named P4R was proposed in [4], where the Llama-2-7b model was used on limited item descriptions to create personalized item profiles for recommendation.

To sum up, LLMs were used for various tasks to enhance recommender system, like LLM-based generative recommendation, retrieval augmented recommendation, LLM-based feature extraction, and conversational recommendation.

3 SCIENTIAREC

This section introduces ScientiaRec for article recommendation, a *Feature-aware Serendipity Matrix Factorization* model that leverages both user-item interactions and content-based features, including article tags and keywords automatically extracted using a LLM.

This model is optimized using a combination of a ranking loss and a serendipity-aware penalty based on feature dissimilarity.

3.1 Feature extraction

To enhance the semantic representation of articles, we enriched their metadata by extracting relevant keywords from their titles and abstracts using the Llama3[1] LLM. Specifically, we employed a local instance of Llama3 (c.f. Table 1). For each document, a prompt was submitted to the model via the ollama interface to generate a list of 10 relevant keywords per article. The prompt used was:

```
Extract exactly 10 distinct and relevant keywords
from the following text. Return them in a single
line, separated only by commas, without any additional
explanation before or after the list. <text> Keywords:
```

To assess the robustness of keyword generation, we also experimented with gemma3[2] and phi-2[3] architectures. Both models produced keyword outputs that were highly similar to those generated by Llama3. Given this consistency and the extended context window offered by Llama3, we adopted it as our default choice for the extraction process.

Table 1: LLaMA3 model configuration.

Model Property	Value
Architecture	llama3
Parameters	8.0B
Context Length	8192 tokens
Embedding Size	4096
Quantization	Q4_0

3.2 Feature-aware Serendipity Matrix Factorization

MF is a widely-used collaborative filtering technique that represents users and items in a shared latent embedding space. Each user u and item i is associated with a dense vector embedding, $\mathbf{u}_u, \mathbf{i}_i \in \mathbb{R}^d$, learned to predict user preferences by modeling interactions through their inner product [13]:

$$\hat{r}_{ui} = \mathbf{u}_u^\top \mathbf{i}_i$$

This captures latent factors explaining user-item interactions, such as tastes and item characteristics, based solely on observed feedback.

To improve personalization and promote serendipitous recommendations, our model extends classical MF by integrating content-based features extracted from item metadata. Specifically, we leverage two types of textual features: user-generated tags and keywords automatically extracted by Llama3 from item titles and abstracts. These features are embedded into the same latent space, allowing the model to capture semantic relationships beyond pure collaborative signals.

[1]https://ollama.com/library/llama3
[2]https://ollama.com/library/gemma3
[3]https://ollama.com/library/phi

For each item i, we compute a *feature embedding* $\mathbf{f}_i \in \mathbb{R}^d$ as the average of its associated feature embeddings:

$$\mathbf{f}_i = \frac{1}{|\mathcal{F}_i|} \sum_{f \in \mathcal{F}_i} \mathbf{e}_f$$

where $\mathcal{F}_i$ is the set of features (tags and keywords) for item i, and $\mathbf{e}_f \in \mathbb{R}^d$ is the embedding vector for feature f.

The final item representation concatenates the original item embedding $\mathbf{i}_i$ with the averaged feature embedding $\mathbf{f}_i$, yielding:

$$\mathbf{x}_i = [\mathbf{i}_i; \mathbf{f}_i] \in \mathbb{R}^{2d}$$

Similarly, user embeddings $\mathbf{u}_u \in \mathbb{R}^{2d}$ are learned to match this augmented item representation.

The predicted preference score is then computed as the dot product in this expanded space:

$$\hat{r}_{ui} = \mathbf{u}_u^\top \mathbf{x}_i = \mathbf{u}_u^\top [\mathbf{i}_i; \mathbf{f}_i]$$

Training Objective. We train the model using Bayesian Personalized Ranking (BPR) [22], a pairwise ranking loss that encourages higher scores for positive items i^+ over negative samples i^-:

$$\mathcal{L}_{\text{BPR}} = -\frac{1}{|S|} \sum_{(u,i^+,i^-) \in S} \log \sigma(\hat{r}_{ui^+} - \hat{r}_{ui^-})$$

To explicitly promote serendipity, we augment this loss with a regularization term based on the Jaccard similarity between item features. This penalizes both positive and negative items that are either too similar or too dissimilar to the user's history, encouraging a balance between relevance and novelty:

$$\mathcal{L} = \mathcal{L}_{\text{BPR}} + \alpha \cdot \mathcal{L}_{\text{Serendipity}}$$

where

$$\mathcal{L}_{\text{Serendipity}} = \frac{1}{|B|} \sum_{(u,i) \in B} \delta_u(i),$$

$$\delta_u(i) = \frac{1}{|\mathcal{I}_u|} \sum_{i' \in \mathcal{I}_u} \max\big(0, |\text{Jaccard}(i, i') - \mu| - \epsilon\big)$$

and B contains both positive and negative candidate items used during training. μ is a target similarity (e.g., 0.30), ϵ is a tolerance (e.g., 0.05), and $\mathcal{I}_u$ denotes the set of items previously interacted with by user u.

Implementation Details. The embeddings are initialized using Xavier uniform initialization and trained using the Adam optimizer. Features are represented via an embedding matrix shared across all tags and keywords. At each training step, the model averages the embeddings of all features associated with an item and concatenates this with the item's own embedding before computing the prediction score. Details about training hyperparameters are presented in Table 2. After 20 epochs of training, our model achieved a BPR loss of 0.0076, a serendipity loss of 0.2426, and a total loss of 0.0319.

4 EXPERIMENTAL STUDY

Based on the RS literature, three main types of evaluation can be distinguished: offline evaluations, user studies, and online evaluations [33]. In this work, we focus on offline evaluation, using a pre-collected dataset that captures users' implicit feedback. This

Table 2: Training hyperparameters.

Hyperparameter	Value
Embedding dimension d	50
Learning rate	0.005
Batch size	512
Epochs	20
Serendipity weight α	0.1
Target similarity μ	0.30
Tolerance ϵ	0.05

approach allows us to assess the performance of our RS without requiring direct user involvement. To gain a deeper understanding of the system's effectiveness, while avoiding the cost and complexity of user-centric evaluations, we complement the offline experiments with an LLM-based evaluation.

We conducted a series of experiments to demonstrate the effectiveness of leveraging LLMs to enrich items representation. Specifically, we compared the top-N recommendation generated by the ScientiaRec approach with those suggested by the same model without LLM-derived features. Additionally, we evaluated our RS against several baseline models based on various evaluation metrics. Finally, we conducted an LLM-based study to evaluate the serendipity of ScientiaRec.

Experiments were run from a laptop equipped of an Intel® Core™ i7-13800H Intel CPU and a NVIDIA RTX A1000 6 GB Laptop GPU. All models were implemented and trained with scripts written in Python 3.12.

4.1 Dataset Description

We use the well know public dataset citeulike-a[4] collected from CiteULike and Google Scholar and contains abstracts, titles, and tags for each article [29]. Its main characteristics are presented in Table 3. We partition the dataset into training data (the earliest 70% papers) and test data (the most recent 30%) for each user.

Table 3: Dataset statistics.

Dataset	Users	Articles	Tags	Citations	User-Article
citeulike-a	5,551	16,980	46,391	44,709	204,987

After the LLM-based enrichment, the number of tags in the citeulike-a dataset reached 99, 696.

4.2 Evaluation Metrics

To assess the effectiveness of using LLMs to enrich items tags and to compare our RS with baseline models, we use two categories of evaluation metrics: accuracy-based, including Precision@N, Recall@N, and F-measure@N, and beyond-accuracy, including Unexpectedness@N, Diversity@N, Novelty@N, and Explainability@N, with $N \in \{5, 10, 15, 20\}$.

[4] https://github.com/js05212/citeulike-a

Precision@N. Measures the proportion of recommended items in the top-N that are relevant.

$$\text{Precision@N} = \frac{|\text{Recommended}_u^N \cap \text{Relevant}_u|}{N}$$

Where $Recommended_u^N$ refers to the list of recommended items @N to the user u and $Relevant_u$ is the list of relevant items to u.

Recall@N. Measures the proportion of relevant items that are successfully recommended.

$$\text{Recall@N} = \frac{|\text{Recommended}_u^N \cap \text{Relevant}_u|}{|\text{Relevant}_u|}$$

F-measure@N. Harmonic mean of *Precision@N* and *Recall@N* that balances the trade-off between precision and recall.

$$\text{F-measure@N} = \frac{2 \cdot \text{Precision@N} \cdot \text{Recall@N}}{\text{Precision@N} + \text{Recall@N}}$$

Unexpectedness@N. Measures the average dissimilarity between recommended items and the user's history. It captures surprising but relevant recommendations.

$$\text{Unexpectedness@N} = \frac{1}{N} \sum_{i \in \text{Recommended}_u^N} \left(1 - \frac{|F_i \cap F_u|}{|F_i \cup F_u|}\right)$$

Where F_i is the feature set of item i and F_u is the union of feature sets from the user's u history.

Diversity@N. Measures the dissimilarity between the recommended items themselves (intra-list dissimilarity). Diversity aims to reduce redundancy within the recommendation list, providing users with a richer and more varied selection of options that better satisfy his/her diverse interests.

$$\text{Diversity@N} = \frac{2}{N(N-1)} \sum_{i<j} \left(1 - \frac{|F_i \cap F_j|}{|F_i \cup F_j|}\right)$$

Novelty@N. Measures how uncommon or rare the recommended items are, based on their popularity. Novelty promotes discovery by encouraging exposure to unpopular content the user is less likely to have encountered on their own, thereby enhancing engagement and exploration.

$$\text{Novelty@N} = \frac{1}{N} \sum_{i \in \text{Recommended}_u^N} -\log_2 \left(\frac{\text{pop}(i) + 1}{|U|}\right)$$

Where $\text{pop}(i)$ is the number of users who have interacted with item i and $|U|$ is the total number of users. The normalized version of the novelty is defined as follow:

$$\text{Norm\_Novelty@N} = \frac{\text{Novelty@N}}{\log_2(|U|)}$$

Explainability@N. Measures how many of the top-N recommended items can be explained to the user by explicitly linking them to his/her past interactions. An item is considered explainable if it shares at least one feature (i.e. tag or keyword) with the items the user has previously consumed. These shared features act as interpretable justifications for why a particular item was recommended.

It helps build trust in the system by offering human-understandable reasons for each suggestion.

$$\text{Explainability@N} = \frac{|\{i \in \text{Recommended}_u^N : F_i \cap F_u \neq \emptyset\}|}{N}$$

Where $F_i \cap F_u \neq \emptyset$ indicates that item i shares at least one feature with the user's u history.

4.3 Baseline models

We compared our RS with other recommendation models. Our goal is to show the effect of using LLMs to enrich items descriptions, and also to compare our recommendation model versus other baselines. The details of the compared methods are listed below:

- **Doc2Vec+kNN:** a content-based RS that suggests research papers by matching a user's profile with top-k similar papers using Doc2Vec [15] abstract embeddings and cosine similarity.
- **CTR:** a simplified variant of the Collaborative Topic Regression model [28], where article topics are extracted via Latent Dirichlet Allocation (LDA) applied to abstracts, and combined with trainable user and item embeddings in a neural framework.
- **CDAE:** an implementation of the Collaborative Denoising AutoEncoder [31] that learns user preferences by combining a corrupted user-item interaction vector with a trainable user embedding in a latent space, reconstructing item scores to predict unseen research papers. The model is trained on implicit feedback from citeulike-a using binary cross-entropy loss, and recommends papers by ranking decoded scores while excluding already read articles.
- **Mult-VAE:** an implementation of the Multinomial Variational Autoencoder for collaborative filtering [16]. The model encodes user interaction vectors into a latent space via a variational autoencoder and reconstructs item probabilities using a decoder. Trained on citeulike-a, the model optimizes a loss combining multinomial log-likelihood and KL divergence for regularization.
- **SGL:** implements the architecture of the Self-supervised Graph Learning model for collaborative filtering [30]. The model constructs a normalized user-item bipartite graph from interaction data, and performs GCN-based embedding propagation across multiple layers. During training, two augmented graph views are generated via random edge dropout, enabling contrastive learning between them. The final loss combines BPR loss for recommendation accuracy and a self-supervised loss to enhance embedding quality.
- **ScientiaRec:** our feature-aware serendipity matrix factorization model for scientific articles recommendation.
- **ScientiaRec$^{\text{w/o LLM}}$:** a second version of our RS ScientiaRec without the use of LLMs to enrich the articles tags.

4.4 Offline Evaluation

We conduct an offline study to compare our RS to other models implemented based on the state of the art architectures. As depicted in Figures 1- 4, ScientiaRec stands out as a balanced model that achieves an effective trade-off between accuracy and beyond-accuracy objectives. It delivers strong performance not only in traditional accuracy metrics like F1@20 (i.e. 0.0259), but also excels in beyond-accuracy metrics including Unexpectedness@20 (i.e. 0.9723), Diversity@20 (i.e. 0.9630), and Explainability@20 (i.e. 0.9824). In particular, from an explainability perspective, ScientiaRec offers the most transparent recommendations, facilitating clear communication with users through statements like: "You are recommended this paper because you previously read articles about Recommender Systems".

When the LLM-derived features are removed, as in the model ScientiaRec$^{\text{w/o LLM}}$, performance drops notably across all metrics, highlighting the importance of the semantic information provided by LLM-generated keywords.

Doc2Vec+kNN demonstrates excellent scores in Novelty and Diversity, however, its extremely low Explainability limits its interpretability. While CTR and SGL achieve the highest Unexpectedness scores (e.g 0.9960 and 0.9949, respectively for Top-20 recommendation) and strong Novelty, yet they suffer from poor accuracy, suggesting a bias toward exploration over relevance. In contrast, Mult-VAE achieves the highest accuracy overall, with a F1@20 reaching 0.0319, while maintaining competitive scores in terms of Unexpectedness and Diversity, making it the most competitive baseline.

4.5 LLM-based Evaluation

Since the concept of serendipity is very complex to assess, we conduct an LLM-based study to measure the serendipity of the list of recommendations generated with our RS ScientiaRec.

4.5.1 Study Design. We selected a sample of 10 users from the citeulike-a dataset, chosen to represent a range of research domains (e.g. recommender systems, social network analysis, bibliometrics, neuroscience, market dynamics, etc.) and asked GPT-4o to evaluate the serendipity of their Top-20 recommended articles generated with our RS ScientiaRec, giving a total of 200 evaluated recommendations.

We asked GPT-4o[5] via the web interface to evaluate how serendipitous a Top-N recommendations are. We used the following prompt:

```
You are simulating an academic researcher with
the following reading history:
1. ID: <text> | Title: <text> ...
Each article consists of a title and you can get
its abstract online (if you need more details
about the article). Based on this history, we
show you a list of recommended articles.
Please evaluate each recommended article from the
perspective of the user. For each recommendation,
provide:
1. Serendipity score: Does it strike a balance
between being relevant and novel — i.e., something
the user wouldn't expect, but might find useful
or interesting?
2. Short explanation (1 sentence)
where score is: Very Low (1), Low (2), Moderate
(3), High (4), Very High (5)
Here is the list of recommended articles:
1. ID: <text> | Title: <text> ...
You can get its abstract online (if you need more
details about the article).
Start by summarizing Key Themes of the user's
reading history items by (domain, examples), like:
```

[5] https://openai.com/index/hello-gpt-4o

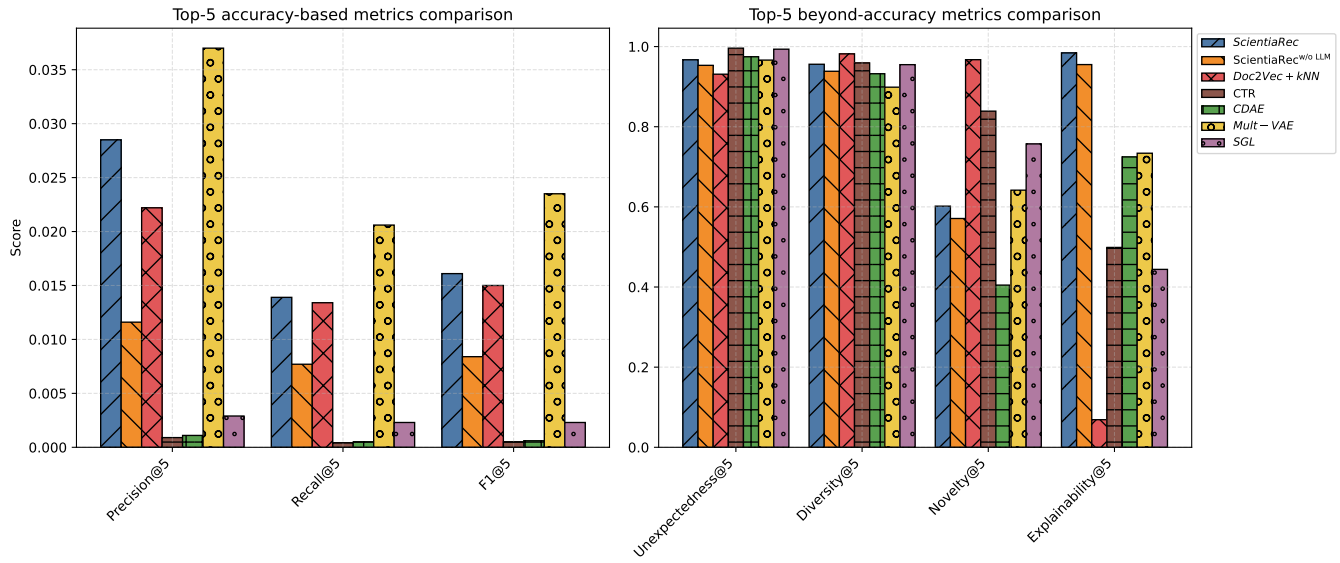

Figure 1: Top-5 recommendation metrics comparison for offline evaluation of different models.

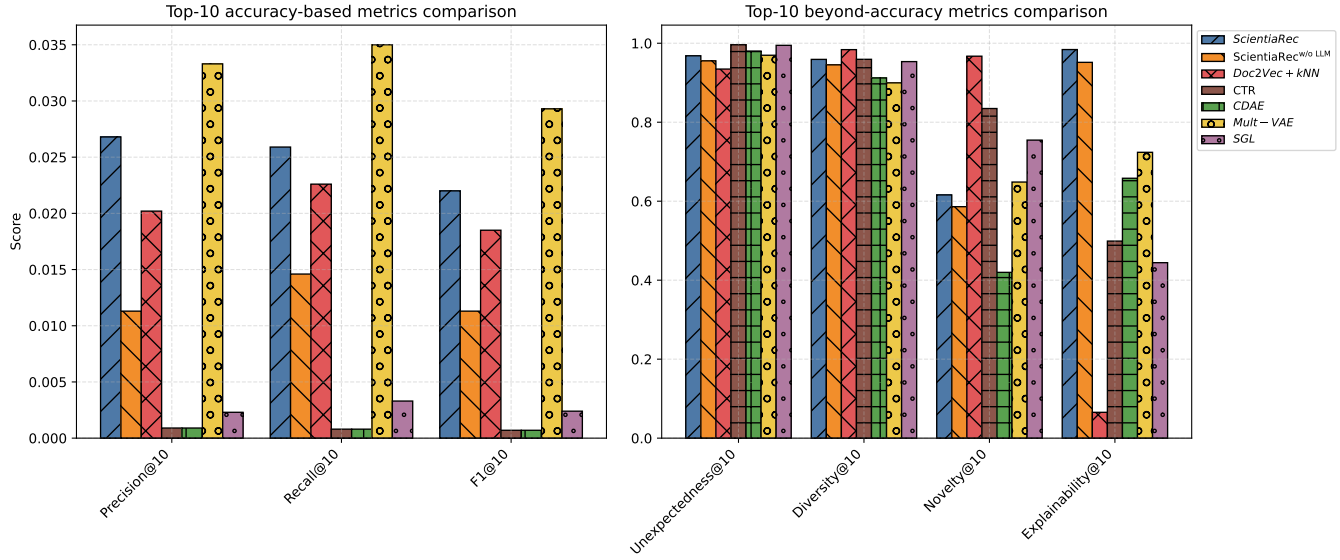

Figure 2: Top-10 recommendation metrics comparison for offline evaluation of different models.

```
<domain> IDs: <text> ...
Next, evaluate the serendipity of the recommendations
— that is, how well they suggest papers that are
not only relevant to the researcher's interests,
but also novel and unexpected.
Format your answer as a table of entries, like:
**Recommendation 1** title, domain
Serendipity: [score, Comments]
...
**Average Serendipity: [score, Comments]
```

4.5.2 Illustrative example. We select a user from the previously sampled users, having the following training history:

```
1. ID: 120 | Title: semantic blogging and decentralized
knowledge management
2. ID: 917 | Title: linked how everything is
connected to everything else and what it means
3. ID: 1836 | Title: the artificial life roots of
artificial intelligence
4. ID: 2489 | Title: evaluating collaborative
filtering recommender systems
5. ID: 2792 | Title: content boosted collaborative
```

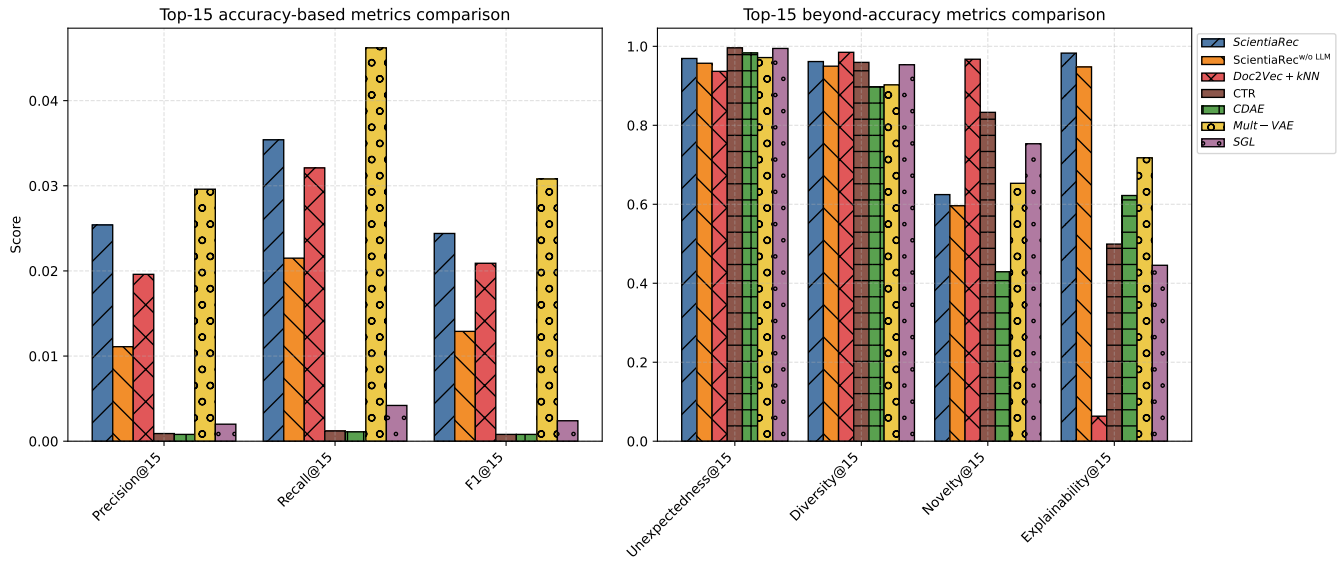

Figure 3: Top-15 recommendation m20trics comparison for offline evaluation of different models.

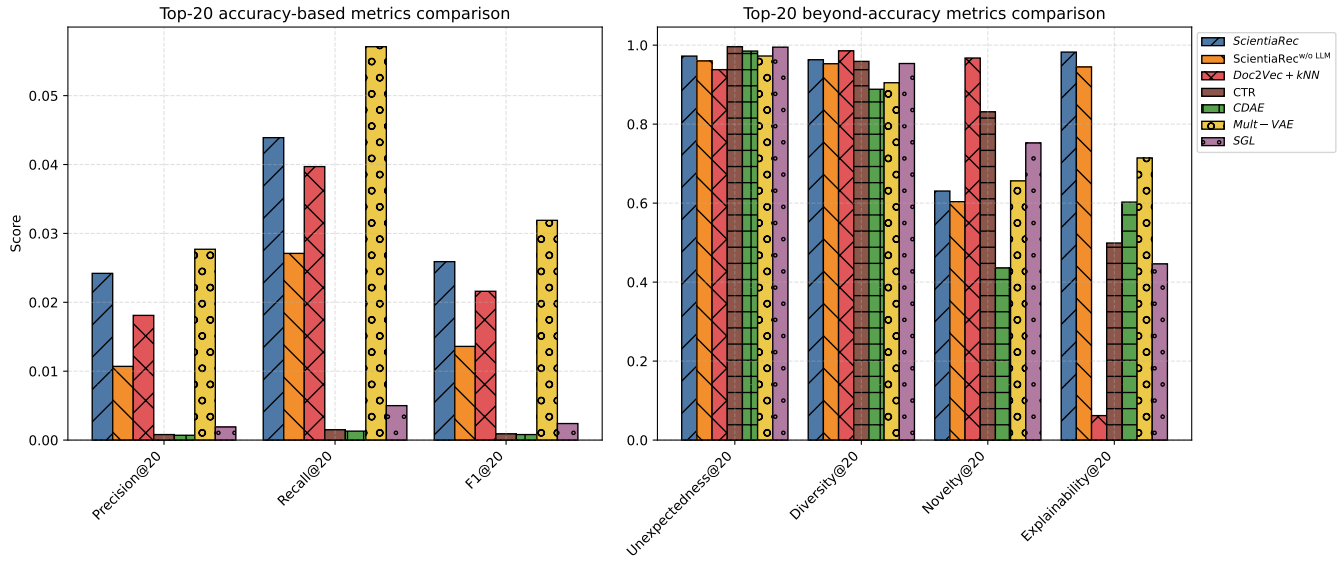

Figure 4: Top-20 recommendation metrics comparison for offline evaluation of different models.

filtering
6. ID: 2793 | Title: recommendation as classification using social and content-based information in recommendation
7. ID: 3981 | Title: the structure of collaborative tagging systems
8. ID: 4050 | Title: the symbol grounding problem
9. ID: 5034 | Title: folksonomy as a complex network
10. ID: 6515 | Title: evolving grounded communication for robots
11. ID: 6598 | Title: feature-based recommendation system
12. ID: 8504 | Title: information retrieval in folksonomies search and ranking
13. ID: 8731 | Title: semiotic dynamics for embodied agents
14. ID: 9542 | Title: the grounding and sharing of symbols
15. ID: 10180 | Title: dont look stupid avoiding pitfalls when recommending research papers
16. ID: 10296 | Title: content-boosted collaborative filtering for improved recommendations
17. ID: 10629 | Title: evolutionary conditions

```
for the emergence of communication in robots
18. ID: 11203 | Title: google news personalization
scalable online collaborative filtering
19. ID: 12205 | Title: collaborative filtering
recommender systems
20. ID: 12206 | Title: content-based recommendation
systems
```

We started by asking the LLM to generate a profile for this user based on his/her reading history. The LLM used the following key themes to define the user's profile:

```
1. Domain: Recommender Systems | IDs: 2489, 2792,
2793, 6598, 10296, 11203, 12205, 12206, 2759,
10180
2. Domain: Tagging & Folksonomies | IDs: 3981,
5034, 8504, 9542
3. Domain: Symbol Grounding & Embodied AI | IDs:
1836, 4050, 6515, 8731, 9542, 10629
4. Domain: Collaborative Filtering | IDs: 2489,
2792, 2794, 4685, 2759
5. Domain: Decentralized Knowledge/Web | IDs: 120,
917, 11205
6. Domain: Evaluation of AI Systems | IDs: 10180,
2759, 4685, 2254
```

Then, we asked the LLM to evaluate the serendipity of the Top-20 recommendations obtained with our RS. The obtained serendipity scores and explanations are detailed in Table 4. The overall serendipity of the Top-20 recommendations was estimated as follows:

```
Average Serendipity: 3.85 (High)
Explanation: The recommendations offer a rich
and diverse expansion from recommender systems
and tagging into social media analysis, incentive
theory, trust epistemology, and tagging dynamics;
many of which are relevant but not obvious, reflecting
genuinely serendipitous discoveries for a user
with strong grounding in symbolic grounding and
folksonomies.
```

4.5.3 Study Results. We detail the results of our LLM-based study in Table 5. This table presents an evaluation of the serendipity of the Top-20 scientific recommendations provided by ScientiaRec to the 10 users sampled for our study, each characterized by different interdisciplinary research domains.

As we can remark, users with broad research interests like User1, User8, and User10, get the recommendations with the highest serendipity score. In these cases, recommendations effectively connect their research domains (extracted from their reading list) to other areas, such as trust epistemology, architecture, or evolutionary psychology. However, User3, with a well-defined and tightly scoped research domain (bibliometrics), encounters moderate serendipity.

As expected, serendipity increases with cognitive distance (when recommending papers that explore new topics or angles), but only when latent coherence is maintained (only if those recommended papers still somehow connect back to the user's research interests). For example, the Top-20 recommendations of User6, interested in search engines and web structure, include collaborative filtering and latent semantic modeling. These two domains are different from classical web search topics, but still connected through shared themes like user behavior, relevance, and information access. Then, ScientiaRec offers something new but not random.

We can also notice that the LLM understood well the notion of serendipity: if a paper is different enough to be surprising, yet still relevant or useful, it hits the sweet spot. So, in its evaluations, high serendipity coincides often with a delicate balance between relevance and unexpectedness. For instance, User5 receives recommendations that combine core neuroscience themes with some surprises like decision variables and synaptic plasticity.

5 CONCLUSION

This paper studied the usefulness of LLMs to enhance scientific papers RS. We first build our RS named ScientiaRec aiming to recommend articles by balancing relevance and serendipity. We used Llama3 to enrich the papers representations by generating keywords from their title and abstract. In addition, we used GPT-4o to evaluate the Top-N recommendations of ScientiaRec in terms of serendipity. Our findings show that LLM-derived keywords can enhance RSs performance. We find out also that LLMs may simulate an academic researcher and evaluate a list of recommendations from a complex notion like serendipity.

Despite the encouraging results presented in this paper, there remains room for improvement. A promising direction for future work is to incorporate article citation information into the recommendation algorithm, rather than relying solely on user–item interactions and item features. Additionally, we plan to evaluate the generalizability of our approach by testing it on other scholarly datasets, such as those from ACM or DBLP. Finally, we also consider leveraging LLMs to address cold-start situations, where user interaction data is limited or unavailable, to improve the applicability of our RS in real-world scenarios.

REFERENCES

[1] Panagiotis Adamopoulos and Alexander Tuzhilin. 2014. On Unexpectedness in Recommender Systems: Or How to Better Expect the Unexpected. *ACM Trans. Intell. Syst. Technol.* 5, 4, Article 54 (2014), 32 pages.

[2] Kurt D. Bollacker, Steve Lawrence, and C. Lee Giles. 1998. CiteSeer: an autonomous Web agent for automatic retrieval and identification of interesting publications. In *Proceedings of the Second International Conference on Autonomous Agents* (Minneapolis, Minnesota, USA) *(AGENTS '98)*. Association for Computing Machinery, New York, NY, USA, 116–123. https://doi.org/10.1145/280765.280786

[3] Pablo Castells, Neil J. Hurley, and Saul Vargas. 2015. *Novelty and Diversity in Recommender Systems.* Springer US, Boston, MA, 881–918.

[4] Junyi Chen and Toyotaro Suzumura. 2024. A Prompting-Based Representation Learning Method for Recommendation with Large Language Models. In *Workshop on Risks, Opportunities, and Evaluation of Generative Models in Recommender Systems (ROEGEN@RecSys 2024)* (Bari, Italy). Association for Computing Machinery, New York, NY, USA.

[5] Li Chen, Yonghua Yang, Ningxia Wang, Keping Yang, and Quan Yuan. 2019. How Serendipity Improves User Satisfaction with Recommendations? A Large-Scale User Evaluation. In *The World Wide Web Conference* (San Francisco, CA, USA) *(WWW '19)*. ACM, New York, NY, USA, 240–250.

[6] Marco de Gemmis, Lops Pasquale, Giovanni Semeraro, and Cataldo Musto. 2015. An investigation on the serendipity problem in recommender systems. *Information Processing & Management* 51, 5 (2015), 695–717.

[7] Yashar Deldjoo, Zhankui He, Julian McAuley, Anton Korikov, Scott Sanner, Arnau Ramisa, René Vidal, Maheswaran Sathiamoorthy, Atoosa Kasirzadeh, and Silvia Milano. 2024. A Review of Modern Recommender Systems Using Generative Models (Gen-RecSys). In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (Barcelona, Spain) *(KDD '24)*. Association for Computing Machinery, New York, NY, USA, 6448–6458.

[8] Driss El Alaoui, Jamal Riffi, Abdelouahed Sabri, Badraddine Aghoutane, Ali Yahyaouy, and Hamid Tairi. 2024. Comparative Study of Filtering Methods for

Table 4: LLM-based evaluation of a user Top-20 recommendations.

Top-20	Title	Theme	Score	Explanation
Rec1	Usage patterns of collaborative tagging systems	Tagging	Moderate	Closely aligned with prior tagging papers; useful, but not especially surprising
Rec2	Shilling recommender systems for fun and profit	RS Security	High	Addresses robustness and manipulation–an underexplored yet important topic for the user
Rec3	GroupLens: an open architecture for collaborative filtering of netnews	CF History	Low	Foundational, but quite expected and potentially already known
Rec4	Amazon.com recommendations: item-to-item collaborative filtering	CF	Moderate	Central to CF literature, slightly predictable but very practically relevant
Rec5	Harvesting social knowledge from folksonomies	Folksonomies	Moderate	Strongly relevant, but an expected continuation of folksonomy theme
Rec6	Review on computational trust and reputation models	Trust Models	High	Broadens scope into trust systems, which intersect with recommendation transparency
Rec7	HT tagging paper taxonomy flickr academic article to read	Tagging Systems	Moderate	Slightly obscure and focused, adds nuance but not unexpected
Rec8	Why we Twitter: understanding microblogging usage and communities	Social Web	Very High	Explores user behavior in a social platform; an unexpected but relevant angle on tagging and recommendation
Rec9	Optimizing web search using social annotations	Web IR	High	Leverages social annotations for IR; strong conceptual link but indirect
Rec10	Tag recommendations in folksonomies	RS with Tags	Moderate	Fits well with past reads, but rather expected
Rec11	Collaborative tagging as a tripartite network	Networked Tagging	High	Introduces a network-theoretic framing; a creative expansion of folksonomy work
Rec12	A taxonomy of incentive patterns	Game Theory / Cooperation	Very High	Diverges into motivational mechanics; novel and intellectually stimulating
Rec13	Toward the next generation of recommender systems	RS Survey	Moderate	Excellent overview, but user likely aware of it
Rec14	Automated tag clustering	Semantic Tagging	High	Enhances tag-based IR and browsing; methodologically fresh
Rec15	Tagging communities: vocabulary evolution	Language Change	High	Focus on how tags evolve over time; subtle but novel angle
Rec16	The complex dynamics of collaborative tagging	Tagging Dynamics	High	Adds depth to folksonomy systems via temporal and behavioral dynamics
Rec17	Why do tagging systems work?	Human Factors in Tagging	Very High	Philosophical and empirical question about tagging behavior; insightful and unexpected
Rec18	Can we trust trust?	Trust & Epistemology	Very High	Theoretical and meta-level reflection on trust in digital systems—deeply novel
Rec19	Clustering methods for collaborative filtering	ML for RS	Moderate	Methodologically relevant, but within core themes
Rec20	Explaining collaborative filtering recommendations	Explainability	High	Strong overlap with interests in recommendation transparency and user experience

Scientific Research Article Recommendations. *Big Data and Cognitive Computing* 8, 12 (2024). https://doi.org/10.3390/bdcc8120190

[9] Yunfan Gao, Tao Sheng, Youlin Xiang, Yun Xiong, Haofen Wang, and Jiawei Zhang. 2023. Chat-REC: Towards Interactive and Explainable LLMs-Augmented Recommender System. arXiv:2303.14524 [cs.IR] https://arxiv.org/abs/2303.14524

[10] Jiwoon Ha, Sang-Wook Kim, Sang-Wook Kim, Christos Faloutsos, and Sunju Park. 2015. An analysis on information diffusion through BlogCast in a blogosphere. *Information Sciences* 290 (2015), 45–62. https://doi.org/10.1016/j.ins.2014.08.042

[11] Leo Iaquinta, Marco de Gemmis, Pasquale Lops, Giovanni Semeraro, Michele Filannino, and Piero Molino. 2008. Introducing Serendipity in a Content-Based Recommender System. In *Proceedings of the 8th International Conference on Hybrid Intelligent Systems (HIS '08)*. IEEE Computer Society, USA, 168–173.

[12] Kirandeep Kaur and Chirag Shah. 2024. Efficient and Responsible Adaptation of Large Language Models for Robust Top-k Recommendations. In *Workshop on Risks, Opportunities, and Evaluation of Generative Models in Recommender Systems (ROEGEN@RecSys 2024)* (Bari, Italy). Association for Computing Machinery, New York, NY, USA.

[13] Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix Factorization Techniques for Recommender Systems. *Computer* 42, 8 (2009), 30–37.

[14] Denis Kotkov, Shuaiqiang Wang, and Jari Veijalainen. 2016. A survey of serendipity in recommender systems. *Knowl.-Based Syst.* 111 (2016), 180–192.

[15] Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32* (Beijing, China) *(ICML'14)*. JMLR.org, II–1188–II–1196.

[16] Dawen Liang, Rahul G Krishnan, Matthew D Hoffman, and Tony Jebara. 2018. Variational autoencoders for collaborative filtering. In *Proceedings of the 2018 world wide web conference* (Lyon, France). 689–698.

[17] Maake Benard Magara, Sunday O. Ojo, and Tranos Zuva. 2018. Towards a Serendipitous Research Paper Recommender System Using Bisociative Information Networks (BisoNets). In *2018 International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD)*. 1–6.

Table 5: LLM-based evaluation study results.

Users	Research domains	Serendipity	Explanation
User1	Recommender Systems, Tagging & Folksonomies, Symbol Grounding & Embodied AI, Collaborative Filtering, Decentralized Knowledge/Web, Evaluation of AI Systems	High	The recommendations offer a rich and diverse expansion from recommender systems and tagging into social media analysis, incentive theory, trust epistemology, and tagging dynamics—many of which are relevant but not obvious, reflecting genuinely serendipitous discoveries for a user with strong grounding in symbolic grounding and folksonomies.
User2	Social Network Analysis & Theories, Complex Networks & Scaling Laws	Moderate	The recommendation list balances core relevance with some refreshing interdisciplinary and theoretical insights. The most serendipitous items draw from outside pure Social Network Analysis, including web science, economics, and network-driven science communication.
User3	Scholarly Publishing & Open Access, Citation Analysis & Bibliometrics	Moderate	While many recommendations are tightly aligned with the user's history, about 6–7 provide meaningful novelty, especially where altmetrics, usage data, access barriers, preprint repositories, and methodological shifts (e.g., PCA) are involved.
User4	Popular Psychology & Behavioral Economics, Children's Literature, Inspirational/Life Advice, Art, Photography & Design, Collaboration & Societal, Finance & Lifestyle	High	Several standout recommendations successfully bridge novelty and relevance. The strongest entries introduced new domains (food, politics, negotiation, fiction) while still resonating with the user's core themes of introspection, lifestyle, and emotional richness.
User5	Cognitive Neuroscience & Sensory Processing, Neuropsychology of Cognition & Culture	High	Overall, this is a very strong and serendipitous set of recommendations. It offers a balanced mix of familiar topics (visual cortex, oscillations) and new cognitive or mechanistic directions (reward, plasticity, decision-making, synaptic timing). The best surprises come from papers that introduce reward prediction, decision variables, and spike-timing-dependent plasticity—topics likely just outside the user's expected scope.
User6	Web Browsing Behavior & Search Engines, Information Retrieval & Clustering, Internet Topology & Network Structure	High	The recommendations strike a strong balance between reinforcing core themes (web search, user modeling, clustering) and introducing novel domains (collaborative filtering, social networks for IR, latent semantic modeling). Several papers offer unexpected yet meaningful extensions of the user's interests.
User7	AI, Machine Learning & Pattern Recognition, Information Retrieval & Search, Recommender Systems & Collaborative Filtering, Web Mining, Folksonomies & Social Tagging	Moderate	This recommendation set is highly relevant, but leans heavily on expected foundational papers. Its serendipity is elevated by articles exploring social tagging, search optimization, and privacy – which effectively bridge distinct domains from the user's interests.
User8	Complex Systems, Networks & Emergence, Web Mining & Semantic Web, Social Cognition & Information Filtering, Information Seeking & Relevance, Meta-Science & Career Advice	High	The recommendation list is well-aligned with the user's interdisciplinary curiosity, blending network science, social cognition, design, and philosophy. High serendipity arises from recommendations that present familiar themes (like networks and relevance) through unfamiliar or cross-domain lenses (e.g., architecture, sociology, design).
User9	MicroRNA Function & Regulation	High	The recommendations strike a strong balance between relevance and novelty. Several entries extend into plant systems, oncology, evolutionary genetics, and transposon biology, offering unexpected yet intellectually rewarding avenues.
User10	Market Dynamics & Agent-Based Modeling, Financial Market Theory & Efficiency, Reinforcement Learning & Game Behavior	High	The recommended set effectively balances relevance (market theory, modeling, decision-making) with novelty (evolutionary psychology, behavioral economics, complex systems). Several suggestions expand the user's view toward human behavior, emergent phenomena, and non-traditional modeling, which are valuable extensions of their current focus.

[18] Christian Matt, Alexander Benlian, Thomas Hess, and Christian Weiß. 2014. Escaping from the Filter Bubble? The Effects of Novelty and Serendipity on Users' Evaluations of Online Recommendations. In *Proceedings of the International Conference on Information Systems*. Auckland, New Zealand, 1–18.

[19] Sheshera Mysore, Andrew Mccallum, and Hamed Zamani. 2023. Large Language Model Augmented Narrative Driven Recommendations. In *Proceedings of the 17th ACM Conference on Recommender Systems* (Singapore, Singapore) *(RecSys '23)*. Association for Computing Machinery, New York, NY, USA, 777–783.

[20] 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* (Seoul, Korea) *(WWW '14)*. Association for Computing Machinery, New York, NY, USA, 677–686. https://doi.org/10.1145/2566486.2568012

[21] Aleksandr V Petrov and Craig Macdonald. 2023. Generative sequential recommendation with GPTRec. In *Proceedings of the workshop on Generative Information Retrieval Gen-IR@SIGIR2023* (Taipei, Taiwan) *(Gen-IR '23)*. https://doi.org/10.1145/3642979.3642995

[22] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian personalized ranking from implicit feedback. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence* (Montreal, Quebec, Canada) *(UAI '09)*. AUAI Press, Arlington, Virginia, USA, 452–461.

[23] Imen Ben Sassi, Priit Järv, and Sadok Ben Yahia. 2024. Does Serendipity Enhance Recommendation Quality? Measuring Accuracy and Beyond-Accuracy Objectives of Serendipitous POI Suggestions. In *ECAI 2024 - 27th European Conference on Artificial Intelligence, 19-24 October 2024, Santiago de Compostela, Spain - Including 13th Conference on Prestigious Applications of Intelligent Systems (PAIS 2024) (Frontiers in Artificial Intelligence and Applications, Vol. 392)*. IOS Press, 3096–3103. https://doi.org/10.3233/FAIA240852

[24] 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 (2015), 91–109.

[25] Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. 2019. BERT4Rec: Sequential Recommendation with Bidirectional Encoder Representations from Transformer. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management* (Beijing, China) *(CIKM '19)*. Association for Computing Machinery, New York, NY, USA, 1441–1450.

[26] Panagiotis Symeonidis, Ludovik Coba, and Markus Zanker. 2019. Counteracting the filter bubble in recommender systems: Novelty-aware matrix factorization. *Intelligenza Artificiale* 13, 1 (2019), 37–47. https://doi.org/10.3233/IA-190017

[27] Saúl Vargas and Pablo Castells. 2011. Rank and Relevance in Novelty and Diversity Metrics for Recommender Systems. In *Proceedings of the Fifth ACM Conference on Recommender Systems* (Chicago, Illinois, USA) *(RecSys '11)*. Association for Computing Machinery, New York, NY, USA, 109–116.

[28] 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* (San Diego, California, USA) *(KDD '11)*. Association for Computing Machinery, New York, NY, USA, 448–456.

[29] Hao Wang, Binyi Chen, and Wu-Jun Li. 2013. Collaborative Topic Regression with Social Regularization for Tag Recommendation. In *IJCAI*.

[30] Jiancan Wu, Xiang Wang, Fuli Feng, Xiangnan He, Liang Chen, Jianxun Lian, and Xing Xie. 2021. Self-supervised graph learning for recommendation. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*. 726–735.

[31] Yao Wu, Christopher DuBois, Alice X. Zheng, and Martin Ester. 2016. Collaborative Denoising Auto-Encoders for Top-N Recommender Systems. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining* (San Francisco, California, USA) *(WSDM '16)*. Association for Computing Machinery, New York, NY, USA, 153–162. https://doi.org/10.1145/2835776.2835837

[32] Zhenrui Yue, Sara Rabhi, Gabriel de Souza Pereira Moreira, Dong Wang, and Even Oldridge. 2023. LlamaRec: Two-Stage Recommendation using Large Language Models for Ranking. In *Proceedings of the workshop on Personalized Generative AI @CIKM2023* (Birmingham, United Kingdom). Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3583780.3615314

[33] Eva Zangerle and Christine Bauer. 2022. Evaluating Recommender Systems: Survey and Framework. *ACM Comput. Surv.* 55, 8, Article 170 (Dec. 2022), 38 pages.

[34] Yuan Cao Zhang, Diarmuid Ó Séaghdha, Daniele Quercia, and Tamas Jambor. 2012. Auralist: Introducing Serendipity into Music Recommendation. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining* (Seattle, Washington, USA) *(WSDM '12)*. Association for Computing Machinery, New York, NY, USA, 13–22. https://doi.org/10.1145/2124295.2124300

[35] 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. *Knowl. Inf. Syst.* 65, 11 (June 2023), 4433–4478. https://doi.org/10.1007/s10115-023-01901-x