

Topic 1. AI-Supported Literature Search



Yong Cao

Postdoc
University of Tübingen
yong.cao@uni-tuebingen.de

Agenda

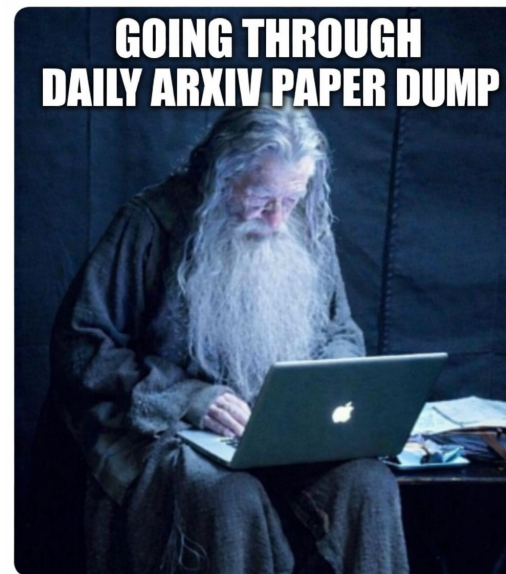
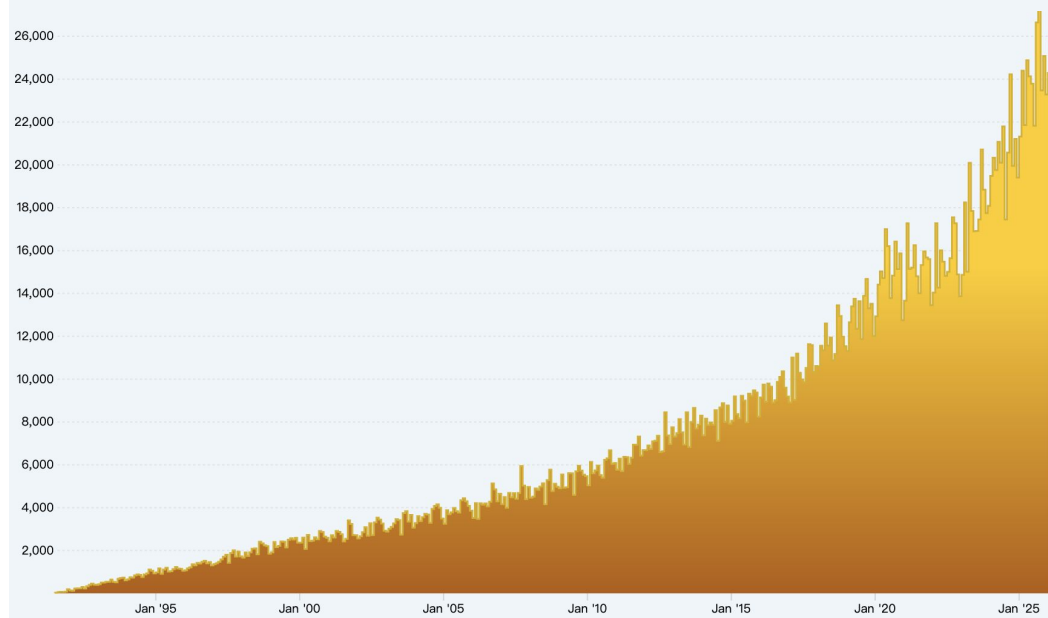
- ❑ Background
- ❑ AI-Enhanced Search System
- ❑ Paper Chat and Scientific QA
- ❑ Graph Based System
- ❑ Recommendation System
- ❑ Takeaway

Agenda

- ❑ **Background**
- ❑ AI-Enhanced Search System
- ❑ Paper Chat and Scientific QA
- ❑ Graph Based System
- ❑ Recommendation System
- ❑ Takeaway

Research Growth

Total number of submissions as of March 28, 2026 = 2,995,510.



[Arxiv Monthly Submission from 1995-2025](#), Arxiv, 2026.3.28.

*The real problem is **not** information overload, it's filter failure.*

— *Clay Shirky*

AI Changes the Way of Scholar's Search

1. Broad Keyword Input

Google 学术搜索

文章 找到约 3,450,000 条结果 (用时0.20秒)

时间不限
2026以来
2025以来
2022以来
自定义范围...

按相关性排序
按日期排序

不限语言
中文网页
简体中文网页

类型不限
评论性文章

包括专利
 包含引用

创建快讯

Exploring the impact of artificial intelligence: Prediction versus judgment
A Agrawal, JS Gans, A Goldfarb - Information Economics and Policy, 2019 - Elsevier
... -making as prediction technology improves. ... prediction impacts the returns to different types of judgment in opposite ways. Hence, not all human judgment will be a complement to AI. ...
☆ 保存 引 引用 被引用次数: 546 相关文章 所有 20 个版本

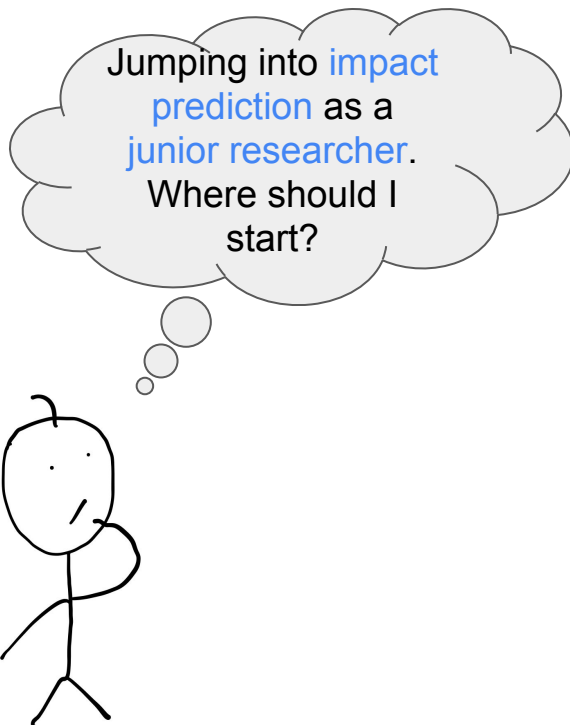
AI-based protein structure prediction in drug discovery: impacts and challenges
M Schaepercl, RA Denny - Journal of Chemical Information and ..., 2022 - ACS Publications
... predict individual domains with high confidence is an impressive step toward solving the protein prediction ... However, we believe that this is just the starting point for AI-based methods to ...
☆ 保存 引 引用 被引用次数: 135 相关文章 所有 3 个版本

AI impact on various domain: an overview
SC Kumain, K Kumain, P Chaudhary - International Journal of ..., 2020 - papers.ssrn.com
... But in strong AI machines can do the calculation and predict the solution for the queries based on Intelligence provided during the training step. In general programming approach, ...
☆ 保存 引 引用 被引用次数: 7 相关文章 所有 3 个版本

Domains and Impact of AI in IoT for Environmental Monitoring and Management: A Review
DS Hammadi, DB Abdullah - 2025 International Conference on ..., 2025 - ieeexplore.ieee.org
... The study in [62] used ML to predict heavy metal levels in soil faster than traditional methods, accelerating the soil rehabilitation process. The prediction of soil metal levels via AI is ...
☆ 保存 引 引用 被引用次数: 2 相关文章 所有 3 个版本

Using artificial intelligence tools in the judicial domain and the evaluation of their impact on the prediction of judgments
S Yassine, M Esghir, O Ibrihich - Procedia Computer Science, 2023 - Elsevier
... artificial intelligence tools and measure their impact on the performance of the judicial system. We study its methods, its challenges, and its potential impacts, ... results for the prediction of ...
☆ 保存 引 引用 被引用次数: 29 相关文章 所有 2 个版本

[PDF] nber.org
[PDF] acs.org
[PDF] ssrn.com
[PDF] researchgate.net
[PDF] sciencedirect.com



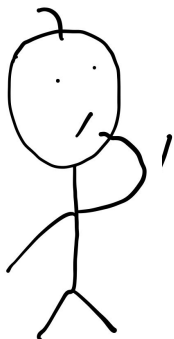
2. Manual Sifting required for relevance

AI Changes the Way of Scholar's Search

1. Detailed Query

2. More accurate papers

Jumping into **impact prediction** as a **junior researcher**.
Where should I start?



Impact prediction in AI domain, focusing on papers published at top AI conferences over the past three years
65 Pro · 2 steps · 20 sources

AI-based scientific impact prediction in recent literature

Recent work shows growing and increasingly sophisticated AI methods for predicting scientific impact, including in AI itself, but only a small subset focuses specifically on top AI conferences.

Main paradigms in scientific-impact prediction

- Impact of ideas, topics, and papers**
 - Knowledge-graph forecasting:** Gu & Krenn build an evolving, 21M-paper semantic-citation graph and use ML to forecast the future growth of research topics with high accuracy (AUC > 0.9), aiming to predict *impact before papers are written* [2].
 - Comprehensive survey:** Xia et al. review impact prediction across four entities (papers, scholars, venues, institutions), cataloging tasks (e.g., citation prediction, h-index forecasting), six feature groups (paper-, author-, venue-, institution-, network-, altmetrics-related), and methods (statistical, ML, deep, graph-based) [3].
 - Paper long-term impact:** Abramo et al. show that integrating early citations, Mendeleev readership, journal impact factor, authorship diversity, and other non-scientific factors improves long-term citation prediction over citation-only models [4]. Content/NLP features and collaboration-network features also enhance accuracy [2].
- Scholar and career impact**
 - FPIN (graph neural network on heterogeneous academic graph)** predicts a scholar's future publication impact after career moves using hierarchical attention and knowledge distillation, outperforming prior models [4].
- Methods, signals, and limitations**

Aspect	Recent insights	Citations
Dominant targets	Future citation counts, topic growth, h-index, institutional/venue impact	[2] [3] [2] [4] [2] [3]
Strong predictors	Early citations, readership, venue quality, collaboration networks	[2] [2] [3]
Advanced models	Deep learning, GNNs, graph-based temporal forecasting	[2] [1] [2] [4] [2] [3]

Ask a follow up...

65 Pro Sources Filter

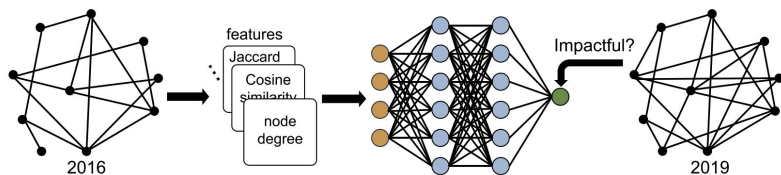
3. Summarization Report

Results

- Forecasting high-impact research topics via machine learning on evolving knowledge graphs**
KEY TAKEAWAY - Machine learning on an evolving knowledge graph can accurately predict the impact of new research ideas, potentially aiding future artificial muses in inspiring new scientific ideas.
2024 · 21 citations · Xuemei Gu et al. · Machine Learning: Science and Its... PDF
- A review of scientific impact prediction: tasks, features and methods**
KEY TAKEAWAY - This paper proposes a new framework for scientific impact prediction, categorizes input features into six groups, and classifies forecasting methods into mathematical statistics, traditional machine learning, deep learning, and graph-based.
LITERATURE REVIEW
2022 · 31 citations · Wanjun Xia et al. · Scientometrics
- Predictive modeling of climate change impacts using Artificial Intelligence: a review for equitable governance and sustainable outcome**
KEY TAKEAWAY - AI-based predictive models can accurately predict climate change impacts, providing valuable insights for scientists, policymakers, and communities.
LITERATURE REVIEW
2025 · 18 citations · K. Ukoba et al. · Environmental Science and Pollut...
- Where is the Next Step? Predicting the Scientific Impact of Research Career**
KEY TAKEAWAY - The Future Publication Impact Prediction Network (FPIN) effectively predicts research scholar's scientific impact using a heterogeneous academic graph and hierarchical attention mechanism.
2025 · 0 citations · Hefu Zhang et al. · IEEE Transactions on Big Data
- Implementing Artificial Intelligence Techniques to Predict Environmental Impacts: Case of Construction Products**
KEY TAKEAWAY - AI techniques can accurately predict environmental impacts of construction products using environmental product declarations, aiding Life Cycle Assessment practitioners and verifiers.
2022 · 44 citations · Anish Koyyamparambath et al. · Sustainability PDF
- Advances in Blast-Induced Impact Prediction—A Review of Machine Learning Applications**
KEY TAKEAWAY - Machine learning methods are better predictors of blast impacts than empirical models, but currently mainly applied in academic research.
LITERATURE REVIEW
2021 · 50 citations · N. Dumakor-Dupey et al. · Minerals PDF

Why AI/LLMs for Literature Search?

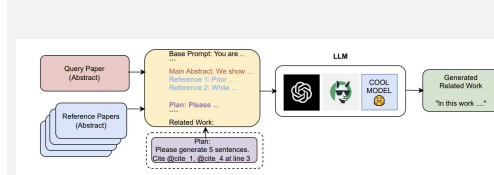
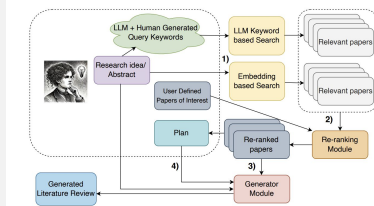
Uncover Emerging Trends



train dataset: 2016 -- 2019; test dataset: holdout data 2016 -- 2019; eval dataset: 2019 --2022

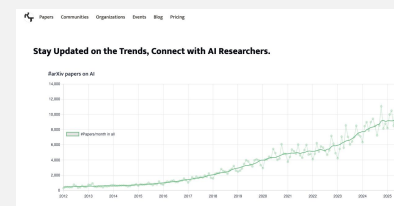
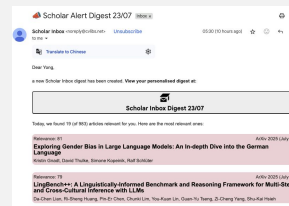
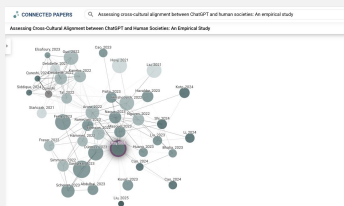
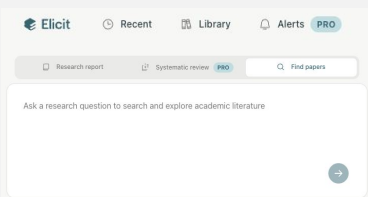
e.g., Forecasting high-impact research topics

Semantic & Relation Understanding



e.g., LLMs for literature review

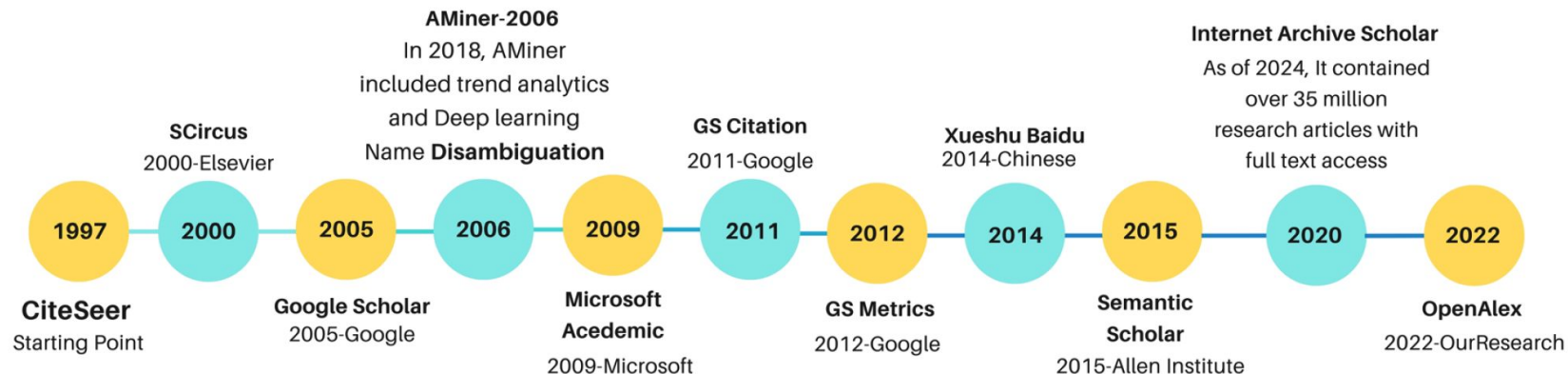
Boost Research Efficiency and Increase Creativity



e.g., Scientific QA, Knowledge Graph Enhanced Search, Semantic Search, personalized Recommendation

Gu et al., [Forecasting high-impact research topics via machine learning on evolving knowledge graphs](#), Machine Learning: Science and Technology, 2025.
 Agarwal et al., [LitLLMs, LLMs for Literature Review: Are we there yet?](#) Transactions on Machine Learning Research, 2025.
[Elicit](#), [Connected Papers](#), [Scholar Inbox](#), [ResearchTrend.ai](#)

Evolution of academic search engines



A shift from simple document retrieval to more intelligent, structured, and open academic ecosystems.

Khalid et al., [Comprehensive review of academic search systems: evolution, analysis, and future research directions](#), Social Network Analysis and Mining, 2025.

Data Source

ResearchTrend.AI



SEMANTIC SCHOLAR



Agenda

- ❑ Background
- ❑ **AI-Enhanced Search System**
- ❑ Paper Chat and Scientific QA
- ❑ Graph Based System
- ❑ Recommendation System
- ❑ Takeaway

AI Enhanced Search System

The screenshot shows the landing page for Undermind, an AI-enhanced search system. At the top left is the Undermind logo (a blue 'U' icon) and the name 'Undermind'. At the top right are links for 'Pricing' and 'Login', and a blue 'Try now' button with a right-pointing arrow. The main heading is 'Condense weeks of research to minutes' in large, bold, black font. Below this is a sub-headline: 'An AI assistant that carefully explores the scientific literature for you. Find exactly what you need, no matter how complex.' There are two buttons: a blue 'Try now' button and a white 'Learn more' button with a light blue border. Below these is a search bar with the placeholder text 'Describe what you're looking for...' and a blue search icon on the right. Underneath the search bar are three example search suggestions in rounded white boxes with light blue borders: 'Experimental evidence of phonon-...', 'Computational models of hippoca...', and 'Laboratory experiments simulating...'

AI Enhanced Search System

The screenshot displays the Undermind web application interface. At the top left, the logo 'Undermind' is visible. The navigation bar includes 'Elicit', 'Recent', 'Library', 'Alerts', and a 'PRO' badge. On the right side of the navigation bar, there are links for 'Pricing', 'Login', and a 'Try now' button. Below the navigation bar, a user profile for 'yongcao2018@gmail.com' is shown with an 'Upgrade' button and a 'Help' dropdown. The main content area features a search bar with tabs for 'Research report', 'Systematic review', and 'Find papers'. Below the search bar, a large text box prompts the user to 'Ask a research question to generate a structured research report'. Underneath this, it says 'Try a couple of free examples to see what this is all about' and provides three example prompts: 'GLP-1R mechanisms', 'Magnesium effects on sleep', and 'Online vs. in-person CBT'. At the bottom, a 'More tools' section contains buttons for 'Upload and extract', 'Summarize concepts', 'Chat with papers', and 'Create an alert' (with a 'PRO' badge).

AI Enhanced Search System

The screenshot displays the Undermind AI search interface. At the top, the Undermind logo is on the left, and navigation links for Pricing, Login, and a Try now button are on the right. Below the navigation, there are links for Elicit, Recent, Library, Alerts, and a PRO badge. A user profile for yongcao2018@gmail.com is visible in the top right. On the left side, there is a sidebar with a large 'Co-res' logo and the text 'An AI assistant you. Find'. The main content area shows a question: 'Can you suggest 3 papers that study LLMs' abilities of generating new research ideas in NLP?'. Below the question, the answer is provided, starting with 'Certainly! Here are three papers that study LLMs' abilities of generating new research ideas in NLP:'. The answer lists three papers with their titles and authors, followed by a brief summary of each paper's content. The interface also includes a 'New Question' button, 'Recent Questions' section, and 'Share' and 'Disclaimer' links.

Undermind Pricing Login Try now →

Elicit Recent Library Alerts PRO Upgrade Help yongcao2018@gmail.com

+ New Question Home Feedback About

Recent Questions

Can you suggest 3 papers that study LLMs' abilities of generating new research ideas in NLP?

Answer Share Disclaimer

Certainly! Here are three papers that study LLMs' abilities of generating new research ideas in NLP:

1. "IdeaBench: Benchmarking Large Language Models for Research Idea Generation" ([Guo et al. 2024¹](#)) ([Guo et al. 2024²](#)) ([Guo et al. 2024³](#))

This paper presents a benchmark for evaluating LLMs' ability to generate research ideas. The authors create a dataset of recent papers from five domains (Computer Science, Economics, Chemistry, Physics, and Medicine) and annotate them with future research ideas. They then evaluate four LLMs (Gemini, Claude-2, GPT-3.5, and GPT-4) on their ability to generate ideas that align with the target papers' ideas. The results show that LLMs can generate research ideas that are novel, relevant, and feasible to a significant extent.

2. "Can Large Language Models Unlock Novel Scientific Research Ideas?" ([Kumar et al. 2024¹](#)) ([Kumar et al. 2024²](#)) ([Kumar et al. 2024³](#))

This paper investigates the potential of LLMs in generating future research ideas across five domains. The authors create a dataset of recent papers from the same five domains as in the previous paper and annotate them with future research ideas. They then evaluate four LLMs (Gemini, Claude-2, GPT-3.5, and GPT-4) on their ability to generate ideas that align with the target papers' ideas. The results show that LLMs can generate research ideas that are novel, relevant, and feasible to a significant extent.

3. "SciPIP: An LLM-based Scientific Paper Idea Proposer" ([Wang et al. 2024](#))

This paper proposes a scientific paper idea proposer (SciPIP) that leverages LLMs to assist researchers in generating new ideas. SciPIP constructs a literature retrieval database and uses LLMs to generate novel and feasible ideas aimed at addressing problems within the given background. The authors demonstrate the effectiveness of SciPIP in generating new research ideas in NLP.

These papers provide insights into the ability of LLMs to generate new research ideas in NLP and highlight the potential of LLMs in accelerating scientific research ([Kumar et al. 2024²](#)).

SEMANTIC SCHOLAR
UNIVERSITY of WASHINGTON

AI Enhanced Search System

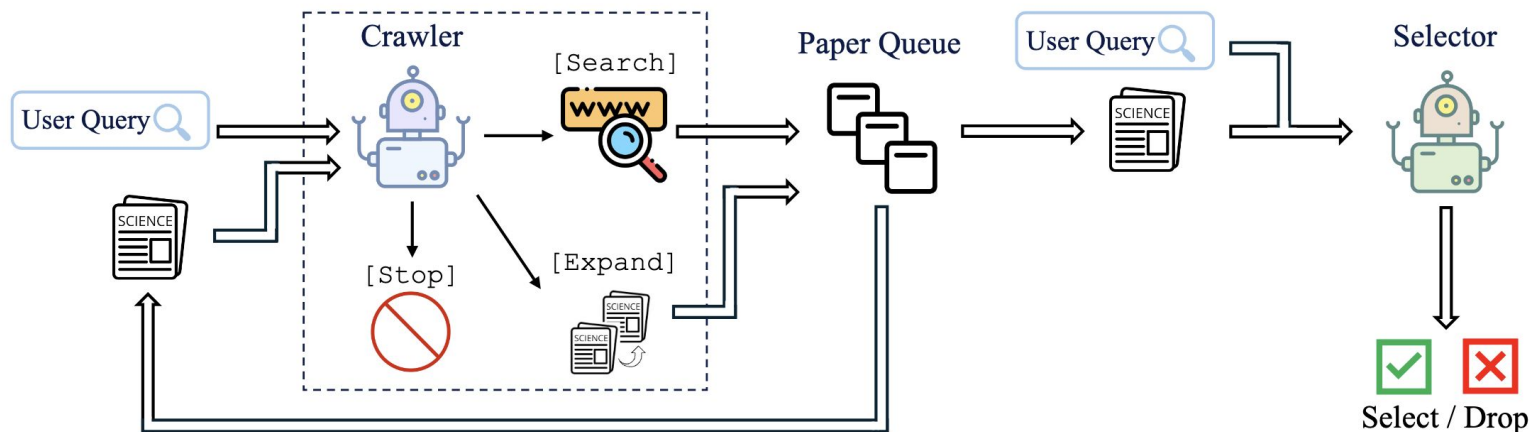
The image displays two overlapping web interfaces. The background interface is 'Undermind', featuring a navigation bar with 'Elicit', 'Recent', 'Library', 'Alerts', and 'PRO'. A sidebar on the left contains the text 'Co res' and 'An AI ass you. Find'. The foreground interface is 'Consensus', showing a search bar with the text 'Ask the research...', a 'Pro' subscription indicator, and a 'Filter' button. Below the search bar are four action buttons: 'Ask a research question', 'Draft an outline', 'Create a table', and 'Try the Consensus Meter'. The main content area of Consensus says 'Research starts here'. At the bottom, it states 'Over 5 million researchers, students, and clinicians trust Consensus'. A 'Sign up' button is visible in the top right of the Consensus interface.

AI Enhanced Search System

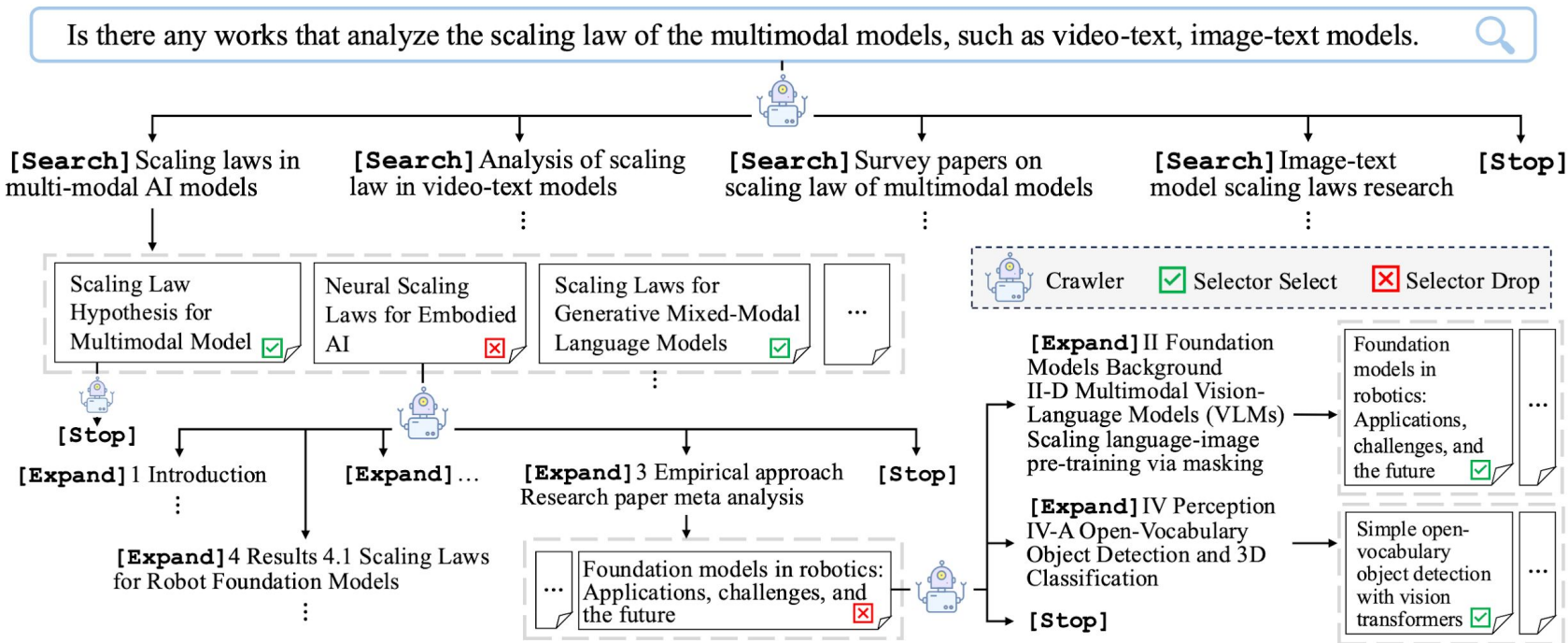
The image displays two overlapping web interfaces. The background interface is 'Undermind', featuring a top navigation bar with 'Pricing', 'Login', and a 'Try now' button. A user profile for 'yongcao2018@gmail.com' is visible. The main content area includes a '+ New Question' button and a 'Recent Question' section. The foreground interface is 'Paperguide', which has a navigation bar with 'Features', 'Deep Research', 'Solutions', and 'Pricing'. The main banner reads 'Your All-in-One AI Research Assistant' and describes the platform as a tool for finding and analyzing research papers. A 'Get Started for Free' button and a 5-star rating from 1000+ researchers are also present.

How AI enhance literature search platforms?

- Mimic human researchers workflows and expand search function.
- Two LLM agents: crawler and selector.



How AI enhance literature search platforms?



Performance

● Imitation Learning + Reinforcement Learning

$$\mathcal{L}_{\text{policy}}(\theta) = \mathbb{E}_{\tau' \sim \pi_{\theta}^{\text{old}}} \left[\min \left(\frac{\pi_{\theta}(a_t|s_t)}{\pi_{\theta}^{\text{old}}(a_t|s_t)} \hat{A}(s_t, a_t), \right. \right. \\ \left. \left. \text{clip} \left(\frac{\pi_{\theta}(a_t|s_t)}{\pi_{\theta}^{\text{old}}(a_t|s_t)}, 1 - \epsilon, 1 + \epsilon \right) \hat{A}(s_t, a_t) \right) \right]$$

$$\mathcal{L}_{\text{value}}(\phi) = \mathbb{E}_{\tau' \sim \pi_{\theta}^{\text{old}}} \left[\max \left(\left(\hat{R}_t - \hat{V}_{\phi}(s_t) \right)^2, \right. \right. \\ \left. \left. \left(\hat{R}_t - \hat{V}_{\phi}^{\text{clip}}(s_t) \right)^2 \right) \right],$$

$$\mathcal{L}_{\text{RL}}(\theta, \phi) = \mathcal{L}_{\text{policy}}(\theta) + \eta \cdot \mathcal{L}_{\text{value}}(\phi)$$

Method	Crawler Recall	Precision	Recall	Recall@100	Recall@50	Recall@20
Google	-	-	-	0.2015	0.1891	0.1568
Google Scholar	-	-	-	0.1130	0.0970	0.0609
Google with GPT-4o	-	-	-	0.2683	0.2450	0.1921
ChatGPT*	-	0.0507	0.3046	-	-	-
GPT-o1	-	0.0413	0.1925	-	-	-
PaSa-GPT-4o	0.7565	0.1457	0.3873	-	-	-
PaSa-7b	0.7931	0.1448	0.4834	0.6947	0.6334	0.5301
PaSa-7b-ensemble	0.8265	0.1410	0.4985	0.7099	0.6386	0.5326

Performance

- Imitation Learning + Reinforcement Learning

Method	Crawler Recall	Precision	Recall	Recall@100	Recall@50	Recall@20
Google	-	-	-	0.2015	0.1891	0.1568
Google Scholar	-	-	-	0.1130	0.0970	0.0609
Google with GPT-4o	-	-	-	0.2683	0.2450	0.1921
ChatGPT*	-	0.0507	0.3046	-	-	-
GPT-o1	-	0.0413	0.1925	-	-	-
PaSa-GPT-4o	0.7565	0.1457	0.3873	-	-	-
PaSa-7b	0.7931	0.1448	0.4834	0.6947	0.6334	0.5301
PaSa-7b-ensemble	0.8265	0.1410	0.4985	0.7099	0.6386	0.5326

Taxonomy-guided Index Construction

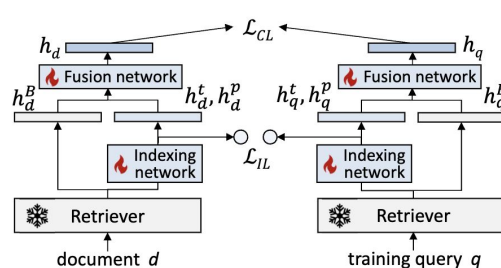
- The Problem: Beyond Surface-Level Text Matching

- TaxoIndex Framework

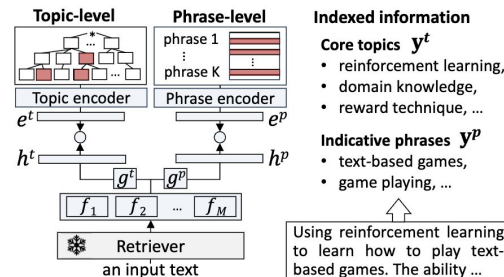
- Step 1: Constructing the Semantic Index

- Step 2: Index-grounded AI Training (Fine-tuning)

- Step 3: Enhanced Retrieval Process



(a) Index-grounded fine-tuning



(b) Index learning with the indexing network

		CSFCube						DORIS-MAE					
		N@5	N@10	M@5	M@10	R@50	R@100	N@5	N@10	M@5	M@10	R@50	R@100
BM25		0.307	0.310	0.088	0.134	0.504	0.635	0.354	0.330	0.079	0.107	0.490	0.669
no Fine-Tuning		0.352	0.337	0.108	0.151	0.524	0.680	0.385	0.360	0.079	0.113	0.551	0.709
FFT		0.372	0.368	0.123	0.169	0.576	0.692	0.408	0.387	0.084	0.122	0.562	0.736
aFT		0.378	0.344	0.119	0.160	0.578	0.696	0.400	0.372	0.080	0.115	0.558	0.714
FFT w/ GRF		0.331	0.317	0.112	0.152	0.561	0.705	0.400	0.379	0.087	0.123	0.586	0.756
FFT w/ ToTER		0.406	0.375	0.135	0.179	0.591	0.710	0.423	0.394	0.091	0.128	0.563	0.736
JTR		0.379	0.352	0.118	0.157	0.598	0.699	0.395	0.380	0.080	0.118	0.548	0.713
TaxoIndex		0.458 ^{†*}	0.417 ^{†*}	0.144 ^{†*}	0.198 ^{†*}	0.633^{†*}	0.741 ^{†*}	0.447 ^{†*}	0.421 ^{†*}	0.104 ^{†*}	0.144 ^{†*}	0.578 [†]	0.756[†]
TaxoIndex ++		0.469^{†*}	0.426^{†*}	0.158^{†*}	0.209^{†*}	0.621^{†*}	0.746^{†*}	0.449^{†*}	0.424^{†*}	0.105^{†*}	0.145^{†*}	0.581[†]	0.751[†]

Deep Research -- ChatGPT / Gemini

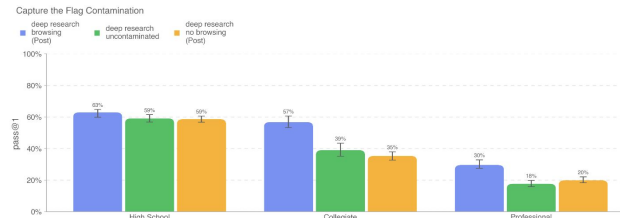
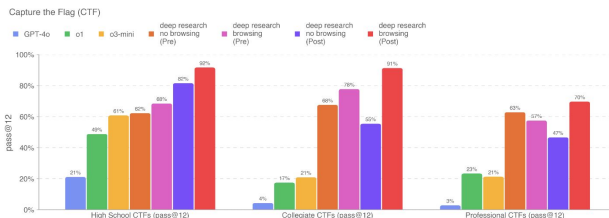
Model data and training

- Browsing datasets
- Graded against the ground truth answers or chain-of-thought model
- Safety datasets from o1 training

Risk and mitigation

- Prompt Injections
- Disallowed Content
- Privacy
- Ability to Run Code
- Bias
- Hallucinations

Overall
Evaluation:
Medium



Planning

Deep Research transforms your prompt into a personalized multi-point research plan

Searching

Deep Research autonomously searches and deeply browses the web to find relevant, up-to-date information

Reasoning

Deep Research shows its thoughts as it reasons over information gathered iteratively and thinks before making its next move

Reporting

Deep Research provides comprehensive custom research reports with more detail and insights, generated in minutes and available as an Audio Overview, saving you hours of time



OpenAI, [Deep Research System Card](#), 2025.

Gemini, [Gemini Deep Research](#), 2026.

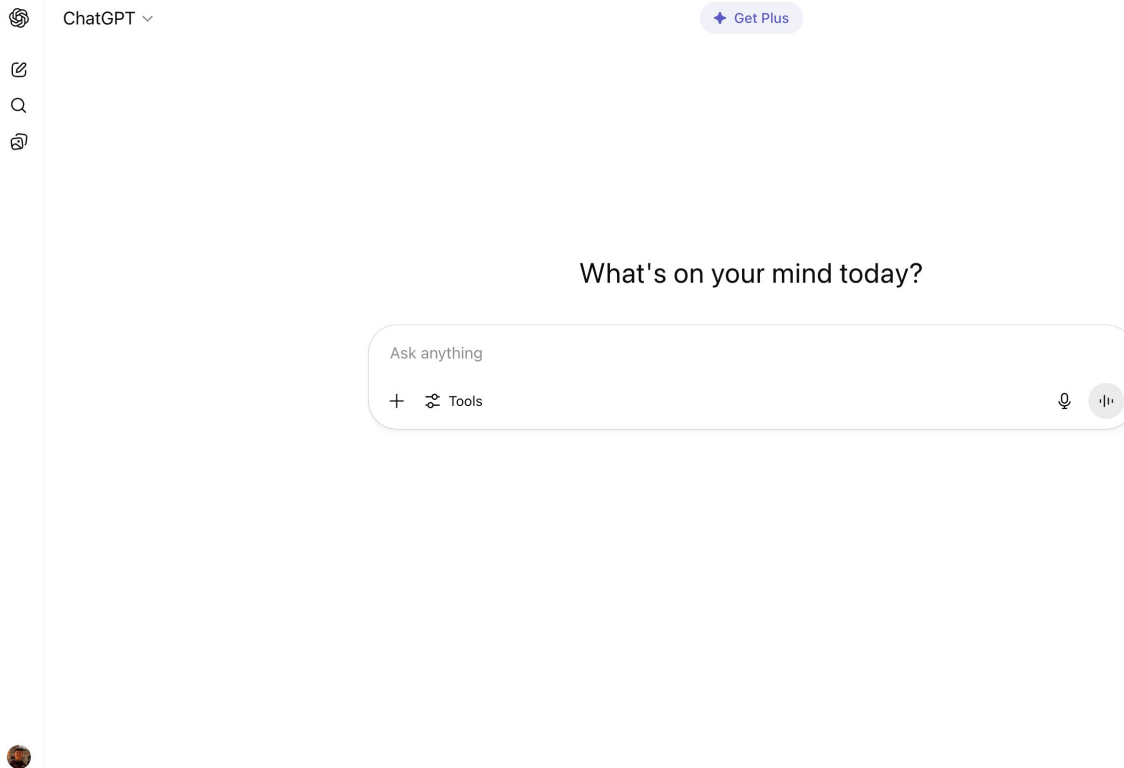
Takeaway – AI Enhanced Search System

- Main Function
 - Optimize scholarly information retrieval
 - Context-aware, semantically rich, and personalized search results
- Key Techniques
 - LLM-based agents
 - Embedding-based Retrieval
 - Personalization
- Challenges
 - Data heterogeneity
 - Limited handling of complex scientific content

Agenda

- ❑ Background
- ❑ AI-Enhanced Search System
- ❑ **Paper Chat and Scientific QA**
- ❑ Graph Based System
- ❑ Recommendation System
- ❑ Takeaway




Paper Chat and Scientific QA



ChatGPT [Get Plus](#)

What's on your mind today?

Ask anything

+  Tools  

Paper Chat and Scientific QA

ChatGPT Get Plus

My Library / Computer science
+ ×

Recent

Reading list 15

Discover

My library

- Computer science 27
- Economics 11
- Machine learning 34
- Health
- Biology 8
- LLMs 10
- Psychology 18
- AI 9
- Sort 34
- Webpages 12
- Crypto
- Meta 6
- Bioinformatics 1
- Sociology 1
- Cyber security 1
- Biotech 1
- Statistics 3
- Philosophy 1
- Knowledge graphs 1
- History 1
- Probability 2
- Physics 3
- Creativity 18
- Astronomy 1
- Classics 1
- Hardware 2
- Uncategorized 2
- Ecology 2
- Cognition 1

New category +

Pro

Trash

Computer science

Collection of papers and articles spanning various subfields of computer science. This library covers topics from foundational concepts to cutting-edge developments, with a particular emphasis on machine learning, artificial intelligence, data analysis, and algorithms.

Authors

Scott Aaronson 1
Tanishq Mathew Abraham 1
Maneesh Agrawala 1
Jessica R. Andrews-Hanna 1
Apple 1
Frank Arute 1
Kunal Arya 1
Ryan Babbush 1
Dave Bacon 1
Max Bain 1

Canvases (5) + Create

Programming GPUs

8 nodes · 11 days ago

Universal media machines

50 nodes · 4 months ago

Turing machines

1 node · last month

New canvas

14 nodes · 3 months ago

Ness

397 nodes · 4 months ago

Items (27) + Add

- Statistical Modeling: The Two Cultures**

Computer science · 2001 · Leo Breiman 👤 3
- The neural basis of loss aversion in decision-making under risk**

Computer science · 2007 · Sabrina M. Tom, Craig R. Fox, Christophe... 👤 0
- The default network and self-generated thought: component processes, d...**

Computer science · 2014 · Jessica R. Andrews-Hanna, Jonathan Small... 👤 3
- Learning to (Learn at Test Time): RNNs with Expressive Hidden States**

Computer science · 2024 · Yu Sun, Xinzhao Li, Karan Dalal, Jiaru X... 👤 0
- Integrated Multi-omics Analysis Using Variational Autoencoders: Appl...**

Computer science · 2019 · Xiaoyu Zhang, Jinggang Zhang, Kai Sun, X... 👤 0
- LMExplainer: Grounding Knowledge and Explaining Language Models**

Computer science · 2023 · Zichen Chen, Jinda Chen, Yuanqian Chen, ... 👤 1
- Scaling Monosemanticity: Extracting Interpretable Features from Claude...**

Computer science · 👤 0
- Mapping the Mind of a Large Language Model**

Computer science · 2024 👤 0
- Mapping inter-city trade networks to maximum entropy models using elec...**

Computer science · 2024 · Cesar I. N. Sampaio Filho, Ridder S. Pi... 👤 3
- Reconstructing the Mind's Eye: fMRI-to-Image with Contrastive Learning...**

53

Paper Chat and Scientific QA



ChatGPT ▾

Get Plus

The screenshot displays the ChatGPT interface with a document titled "LLMBench: Large Language Model Compression Benchmark" open. The interface is divided into three main sections: Sources, Chat, and Studio.

- Sources:** A sidebar on the left lists various categories under "My Library". The "Computer science" category is selected. A search bar contains "LLMBench.pdf", which is listed as a source with a checkmark.
- Chat:** The main content area shows the document title and a summary. The text reads: "The document introduces LLMBench, a novel benchmark designed to systematically evaluate and compare various large language model (LLM) compression techniques. It addresses the current challenges in LLM deployment, such as high computational demands and limited, inconsistent evaluation standards across existing research. LLMBench establishes six distinct evaluation tracks, including compression performance, generalization ability, training and inference consumption, hardware acceleration, and model trustworthiness, to offer a comprehensive analysis. By benchmarking mainstream sparsification and quantization methods across diverse models, datasets, and deployment platforms, the paper aims to provide actionable insights for the design and practical application of more efficient LLM compression algorithms. Ultimately, it seeks to standardize evaluation protocols and guide future research toward making LLMs more accessible for real-world use." Below the text is a chat input field with the prompt: "What are the most significant challenges hindering the practical deployment of large language models?"
- Studio:** A sidebar on the right contains an "Audio Overview" section with a "Generate" button, a "Deep Dive conversation" section with a "Generate" button, and a "Notes" section with an "Add note" button and options for "Study guide", "Briefing doc", "FAQ", and "Timeline".

At the bottom of the interface, there is a footer with the text: "NotebookLM can be inaccurate; please double check its responses." Below this, there are two document thumbnails: "Mapping inter-city trade networks to maximum entropy models using elec..." and "Reconstructing the Mind's Eye: fMRI-to-image with Contrastive Learning..."

Paper Chat and Scientific QA



ChatGPT ▾

Get Plus

The screenshot displays a web browser window with the enago Read interface. The main article is titled "Light-Induced Self-Oscillations and Spoiling of the Bragg Resonance Due to Nonlinear Optical Propagation in Helicoidal Cholesteric Liquid Crystals" by Ashot H. Gevorgyan and Francesco Simoni. The article includes an abstract, a citation, and keywords. On the right side, a chat interface is visible, showing a question about the research results and a response from the AI.

Computer science

LLMCBench: Large Language Model Compression Benchmark

enago Read

Search

Set Preferred Chat Language: English (en)

the light wavelength is close to the Bragg resonance. The study demonstrates that self-induced oscillations take place at lower intensities on the red side edge of the Bragg resonance while stable values of transmittivity are still observed on the blue side edge. A further increase in intensity leads to oscillations of lower amplitude on the blue side and an irregular behavior of the transmission on the red side. At higher intensities, the Bragg resonance disappears and transmission becomes unstable for any light wavelength. A simple phenomenological model is proposed to account for the onset of the oscillations and the asymmetry of the behavior at the opposite side of the Bragg resonance. The study also points out that the static electric field is a driving parameter to switch from stable to oscillatory to irregular behavior in the transmittivity at a given light wavelength. The research received no external funding and the authors declare no conflict of interest. Data are available from the authors.

08:00 pm

The research reports the results of a study on the nonlinear light propagation of a beam traveling along the helix...

Ask your question...

check for updates

Citation: Gevorgyan, A.H.; Simoni, F. Light-Induced Self-Oscillations and Spoiling of the Bragg Resonance Due to Nonlinear Optical Propagation in Helicoidal Cholesteric Liquid Crystals. *Photonics* **2022**, *9*, 881. <https://doi.org/10.3390/>

Keywords: helicoidal liquid crystals; nonlinear optics; pitch tuning; optical instabilities

1 / 12 121 %

Paper Chat and Scientific QA



ChatGPT ▾

Get Plus

My Library / Computer science
Computer science

Recent

Reading list

Discover

My library

- Computer science
- Economics
- Machine learning
- Health
- Biology
- LLMs
- Psychology
- AI
- Sort
- Webpages
- Crypto
- Meta
- Bioinformatics
- Sociology
- Cyber security
- Biotech
- Statistics
- Philosophy
- Knowledge graphs
- History
- Probability
- Physics
- Creativity
- Astronomy
- Classics
- Hardware
- Uncategorized
- Ecology
- Cognition

New category +

LLMCBench: Large Language Model Compression Benchmark

enago Read

Sources

+ Add

Select all source

LLMCBench

Explainpaper

100%

1 / 13

Paper Explain Chat

LLMCBench: Benchmarking Large Language Model Compression for Efficient Deployment

Ge Yang¹, Changyi He¹, Jinyang Guo¹, Jianyu Wu¹, Yifu Ding¹, Aishan Liu¹, Haotang Qin², Pengliang Ji³, Xianglong Liu¹

¹ Beihang University ² ETH Zurich ³ Carnegie Mellon University

Abstract

Although large language models (LLMs) have demonstrated their strong intelligence ability, the high demand for computation and storage hinders their practical application. To this end, many model compression techniques are proposed to increase the efficiency of LLMs. However, current researches only validate their methods on limited models, datasets, metrics, etc, and still lack a comprehensive evaluation under more general scenarios. So it is still a question of which model compression approach we should use under a specific case. To mitigate this gap, we present the Large Language Model Compression Benchmark (LLMCBench), a rigorously designed benchmark with an in-depth analysis for LLM compression algorithms. We first analyze the actual model production requirements and carefully design evaluation tracks and metrics. Then, we conduct extensive experiments and comparison using multiple mainstream LLM compression approaches. Finally, we perform an in-depth analysis based on the evaluation and provide useful insight for LLM compression design. We hope our LLMCBench can contribute insightful suggestions for LLM compression algorithm design and serve as a foundation for future research. Our code is available at <https://github.com/AboveParadise/LLMCBench>.

1 Introduction

Recently, large language models (LLMs) have attracted increasing attention because of their strong intelligence ability. While it achieves excellent performance, the huge computation and storage burden hinders the practical usage of these LLMs. To solve this problem, many model compression

31 Oct 2024

Citation

Light

Spillo

to No

Helic

Cry

htps:

Upgrade

Dashboard

Upload

To mitigate this gap, we present the Large Language Model Compression Benchmark (LLMCBench), a rigorously designed benchmark with an in-depth analysis for LLM compression algorithms.

UNDERGRAD

To address the challenges in evaluating large language model (LLM) compression methods, the authors introduce the Large Language Model Compression Benchmark (LLMCBench). This benchmark is carefully designed to provide a thorough analysis of different LLM compression algorithms, allowing for better comparisons and insights into their effectiveness. By establishing this benchmark, the authors aim to improve the understanding and development of efficient LLM compression techniques.

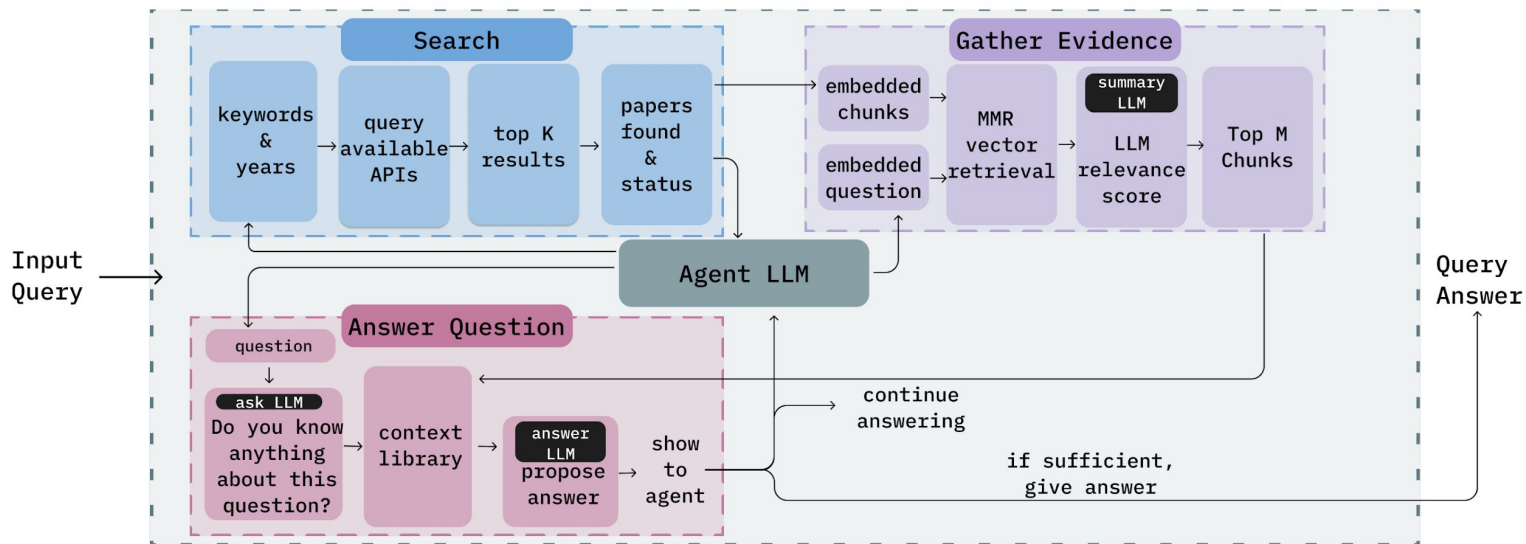
Ask a follow-up

PaperQA – An Agent-Based RAG System

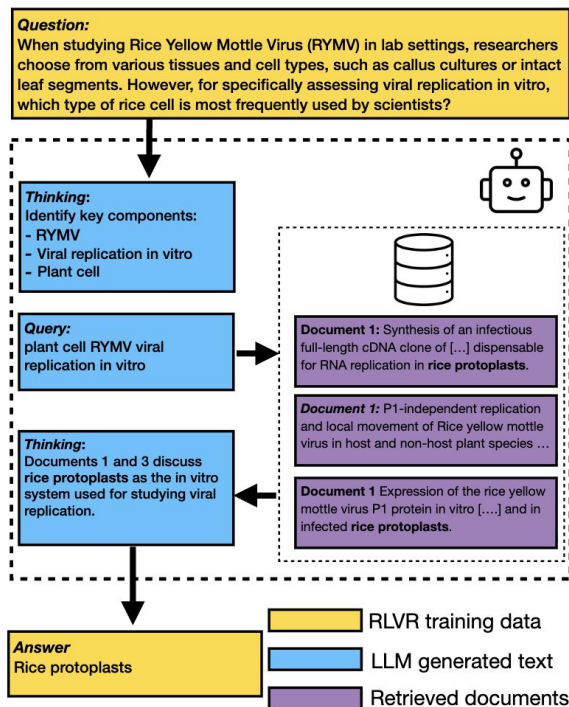
- Aims to address the hallucinations and knowledge LLM update.

- Key Components:

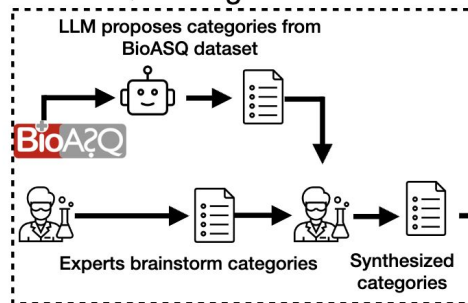
- Search, Gather Evidence, Answer Question



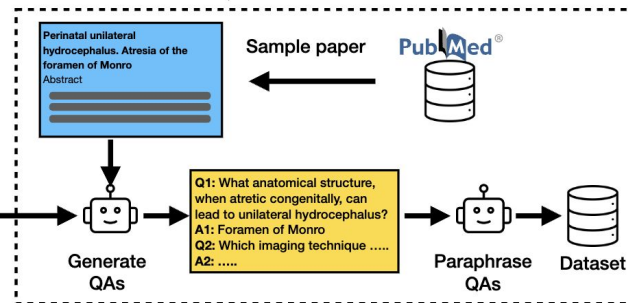
Reinforcement Learning Helps



Define QA Categories

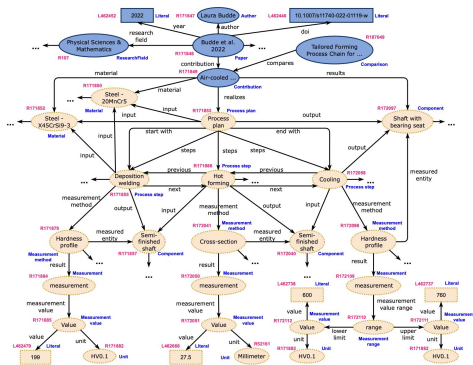


Generate new QAs



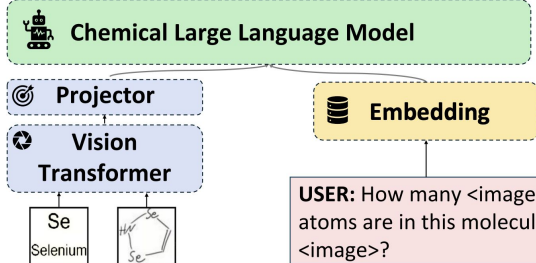
	PaperSearchQA	BioASQ
Qwen2.5-3b-Instruct		
Direct	16.7	15.8
CoT	20.3	16.5
RAG	32.0	30.0
Search-o1	30.8	29.4
PaperQA2	32.4	33.1
SearchR1	41.6	35.5
Qwen2.5-7b-Instruct		
Direct	27.5	24.9
CoT	29.7	23.4
RAG	36.5	29.7
Search-o1	36.5	31.5
PaperQA2	37.1	32.8
SearchR1	51.0	44.8

Mostly focus on...

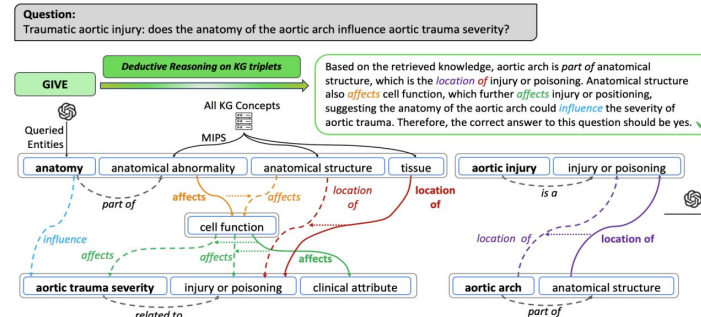


KG Integration

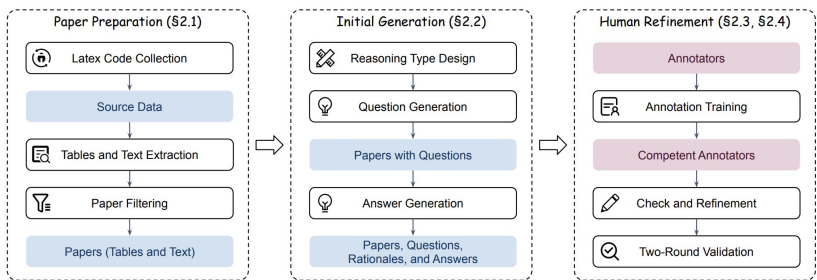
ASSISTANT: I think the chemical structure in this image is $C1=C[Se]N[Se]1$. So there were two Selenium atoms.



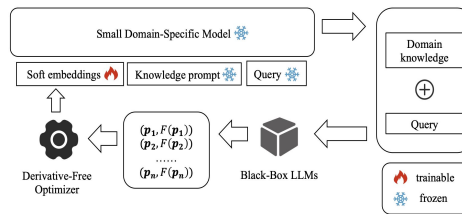
Multi-domal QA



Complex Reasoning



Benchmarking



Domain knowledge

Auer et al., [The SciQA Scientific Question Answering Benchmark for Scholarly Knowledge](#), Scientific Reports, 2023.

Li et al., [ChemVLM: Exploring the Power of Multimodal Large Language Models in Chemistry Area](#), AAAI, 2025.

Li et al., [BLADE: Enhancing Black-box Large Language Models with Small Domain-Specific Models](#), AAAI, 2025.

Zhang et al., [SCITAT: A Question Answering Benchmark for Scientific Tables and Text Covering Diverse Reasoning Types](#), ACL-findings, 2025.

He et al., [GIVE: Structured Reasoning of Large Language Models with Knowledge-Graph-Inspired Veracity Extrapolation](#), ICML, 2025.

Takeaway – Paper Chat and Scientific QA

- Main Function
 - Interactive Q&A with research papers
- Key techniques
 - LLM-based agents
 - PDF Parsing & Contextual Chunking
 - Summarization & Key Point Extraction
 - Multi-turn Dialogue Tracking
- Challenges
 - PDF Quantity & Size Limits
 - Domain Knowledge
 - Reliability & Explainability

Agenda

- ❑ Background
- ❑ AI-Enhanced Search System
- ❑ Paper Chat and Scientific QA
- ❑ **Graph Based System**
- ❑ Recommendation System
- ❑ Takeaway

Graph Based System

CONNECTED PAPERS

Share Follow About Pricing Log in

Explore connected papers in a visual graph

To start, enter a paper identifier

Search by keywords, paper title, DOI or another identifier [Build a graph](#)

You can try:



Or start with one of our example graphs



Graph Based System

CONNECTED PAPERS

Share Follow About Pricing Log in

ScholarGPS
POWERING SCHOLARLY ANALYTICS

All Search for any scholar, Specialty, Discipline, Field, or Institution

SEARCH

Log in | Register

Try: Anil K.Jain | Michigan State University | Computer Science | Pattern recognition

Scholar Profiles Institutional Profiles Highly Ranked Scholars Institutional Rankings Highly Cited Publications More

Empowering scholars to accelerate research and collaboration

Profiles and Rankings of all Scholars, Institutions, Countries, Fields, Disciplines, and Specialties

What is ScholarGPS

Key Features and Benefits

ScholarGPS News and Analyses

Quick Tips

ScholarGPS® gives unprecedented access to:



Scholar Profiles

Access the profiles of over 30 million academic, industrial, and government scholars from 120,000 institutions in over 200 countries. Claim and update your profile — for free!

[More](#)



Institutional Profiles

Access analytics for each of the 120,000 institutions (24,353 academic institutions) included in the ScholarGPS® database.

[More](#)



Field and Discipline Profiles

Access descriptions of, and detailed analytics for broad 14 Fields and 177 Disciplines. [More](#)



Specialty Profiles

Access descriptions of, and detailed analytics for over 350,000 distinct Specialties. [More](#)

Graph Based System

CONNECTED PAPERS

Share Follow About Pricing Log in

ScholarGPS
POWERING SCHOLARLY ANALYTICS

All Search for any scholar, Specialty, Discipline, Field, or Institution

SEARCH

Try: Anil K. Jain | Michigan State University | Computer Science | Pattern recognition

Log in | Register

What captures our attention may not always guide us to the best path, while what we fail to notice often makes the difference.

Highly Cited Papers

The deeper, more inspirational, and timely signals often hide beneath the surface, where intriguing ideas emerge and groundbreaking research begins.

CiteSpace
brings them to light.

#14 foreign influence
#9 press coverage
#19 mass-media
#7 television-news format
#15 global analysis
#42 social control
#40 warfare
#10 military commission
#8 future conventional battlefield
#3 domestic terrorist attack
#16
#49 war trauma survivor
#0 biological weapon
#20 latin

Graph Based System

CONNECTED PAPERS

Share Follow About Pricing Log in

ScholarGPS[®]
POWERING SCHOLARLY ANALYTICS

All Search for any scholar, Specialty, Discipline, Field, or Institution

SEARCH

Try: Anil K. Jain | Michigan State

Log in | Register

CiteSpace

What captures our attention may

#14 foreign

#9 press coverage

#19 mass-

#7 television-news format

#15 gl

Sci2 Tool
A Tool for Science of Science Research & Practice

Home Download Documentation Ask An Expert Testimonials Developers

This site is deprecated, please download the latest version of Sci2 on [GitHub](#).

The Science of Science (Sci2) Tool is a modular toolset specifically designed for the study of science. It supports the temporal, geospatial, topical, and network analysis and visualization of scholarly datasets at the micro (individual), meso (local), and macro (global) levels.

[Registration required.](#)

Download Sci² Tool

News

2018

- Jan 31, The [Sci2 \(Science of Science\) Tool v1.3](#) release provides support for Java 1.9 and removes deprecated plugins. [Release Notes](#)

[More news](#)

An Emergent Mosaic of Wikipedian Activity

2 / 4 Start Stop (3)

Have a question? Ask an Expert!

,353
abase.

Graph Based System

CONNECTED PAPERS

Share Follow About Pricing Log in

ScholarGPS®
POWERING SCHOLARLY ANALYTICS

All Search for any scholar, Specialty, Discipline, Field, or Institution

SEARCH

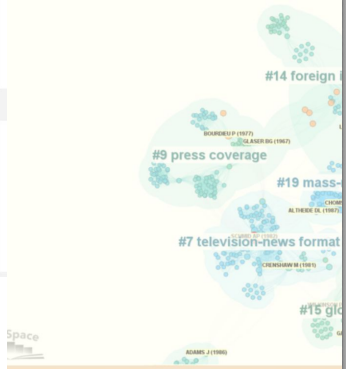
Log in Register

Try: Anil K. Jain | Michigan State



What captures our attention may

S



Search My library ORKG

English



Find research you are *actually* looking for

Ask your question...



ORKG Ask is a scholarly search and exploration system powered by **Vector Search, Large Language Models and Knowledge Graphs**. [Learn more.](#)

77,368,538
Items with abstracts

Getting started

- How does the availability of green spaces impact community health and well-being?
- What role does storytelling play in shaping collective memory and identity?
- What impact does access to quality early childhood education have on lifelong outcomes?
- What are the effects of sleep deprivation on cognitive performance and overall health?

ORKG Ask is brought to you by



This site is depre

The Science of S
specifically desig
the temporal, ge
visualization of s
meso (local), and
[Registration req](#)



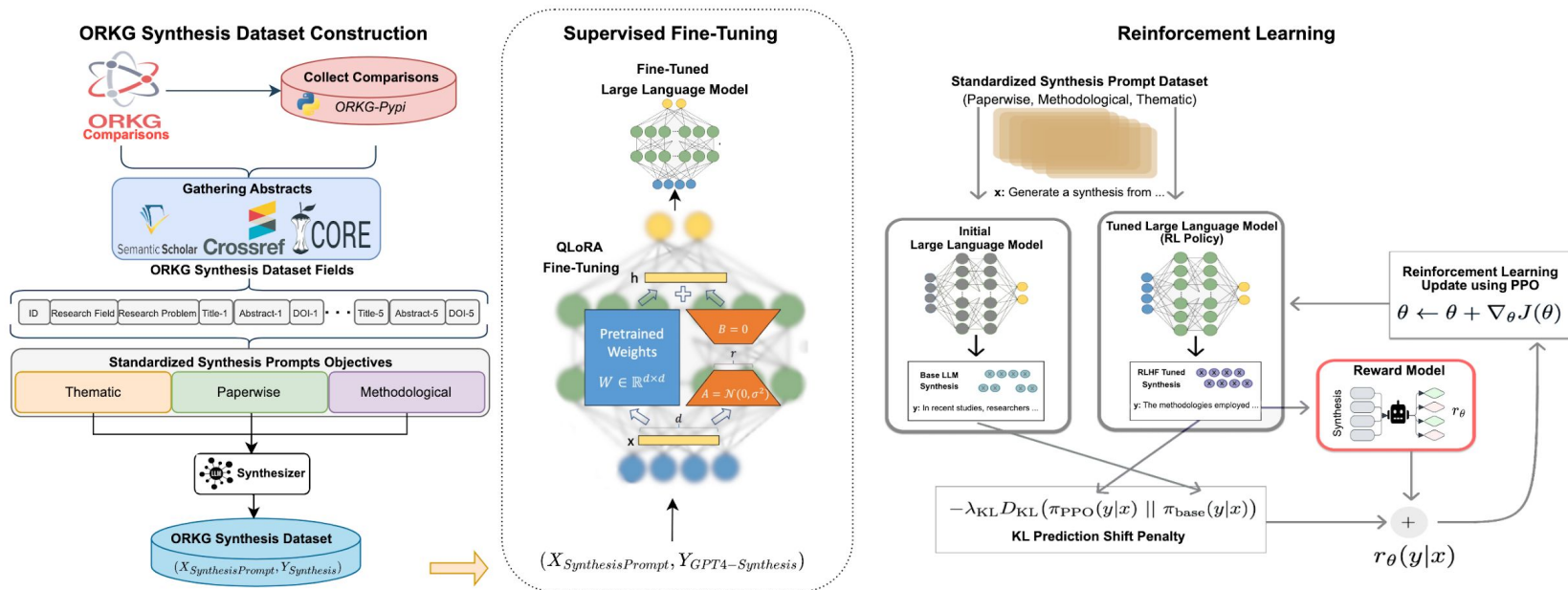
News

2018

- Jan 31, TI
release pi
deprecat
[Release N](#)

[More news](#)

Graph Based System



Performance and Conclusion

- SFT+RLAIF performs the best.

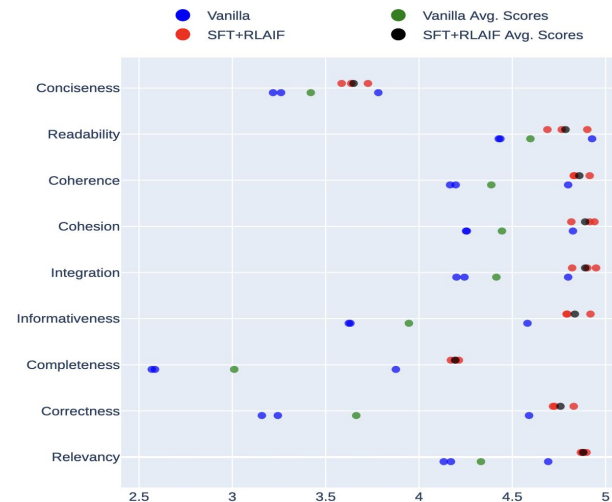
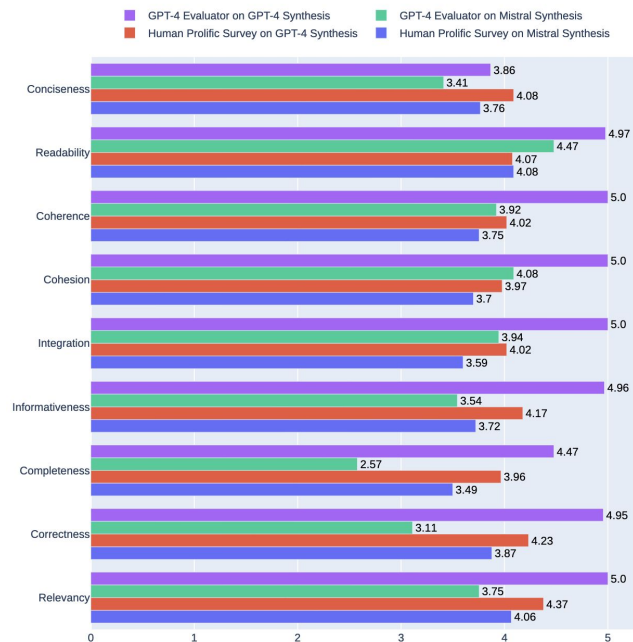
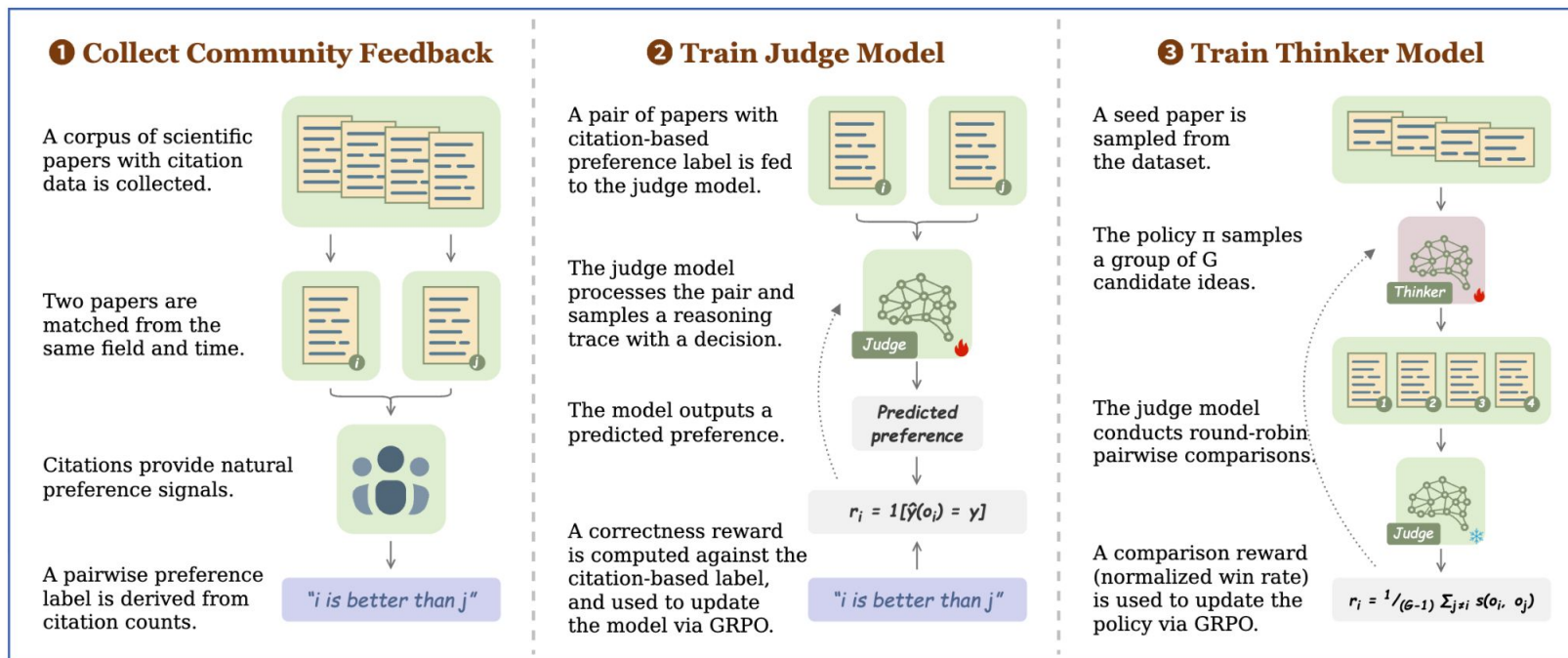


Figure 3: Consistency comparison of the GPT-4 evaluator between the *Vanilla* and *SFT+RLAIF (w/ GPT-4 Features)* models, assessed through three evaluations on the test set.

AI Can Learn Scientific Taste?

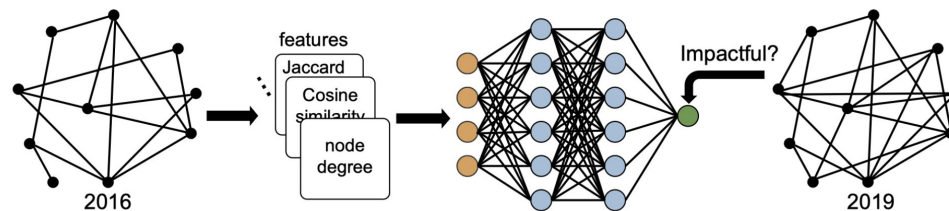
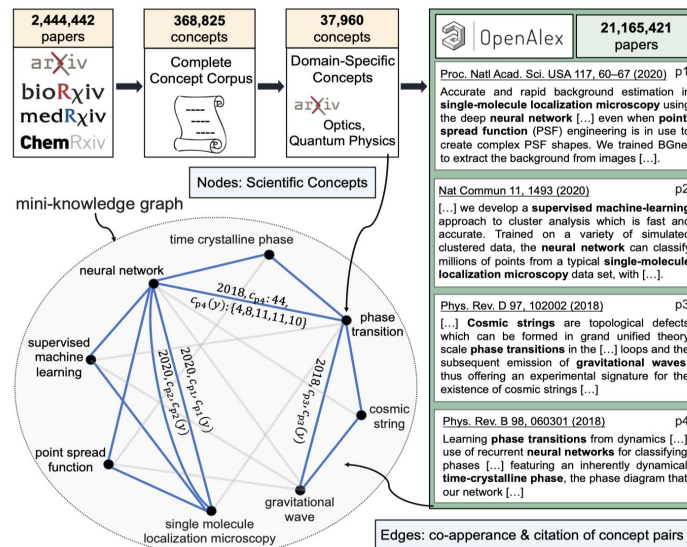


AI Can Learn Scientific Taste?

Model	CS	Math	Physics	Others	Avg.
<i>Open-source Models</i>					
Qwen3-4B-Instruct	66.5	65.6	54.8	57.1	60.3
SciJUDGE-Qwen3-4B	78.6 (+12.1)	74.6 (+9.0)	71.2 (+16.4)	79.8 (+22.7)	75.3 (+15.0)
Qwen3-30B-A3B-Instruct	73.8	70.5	59.4	65.5	66.3
SciJUDGE-Qwen3-30B	83.5 (+9.7)	78.7 (+8.2)	78.7 (+19.2)	82.3 (+16.8)	80.6 (+14.3)
Qwen2.5-1.5B-Instruct	6.3	10.7	6.0	6.7	7.0
SciJUDGE-Qwen2.5-1.5B	72.3 (+66.0)	73.0 (+62.3)	69.4 (+63.4)	77.3 (+70.6)	72.1 (+65.1)
Qwen2.5-3B-Instruct	16.5	36.9	23.8	21.0	23.5
SciJUDGE-Qwen2.5-3B	76.2 (+59.7)	76.2 (+39.3)	66.2 (+42.3)	81.5 (+60.5)	73.2 (+49.7)
Qwen2.5-7B-Instruct	57.3	37.7	37.0	51.3	45.2
SciJUDGE-Qwen2.5-7B	83.0 (+25.7)	68.8 (+31.1)	71.5 (+34.5)	87.4 (+36.1)	76.9 (+31.7)
Qwen2.5-14B-Instruct	64.1	63.9	54.5	56.3	59.1
SciJUDGE-Qwen2.5-14B	87.9 (+23.8)	78.7 (+14.8)	74.4 (+19.9)	84.9 (+28.6)	80.6 (+21.6)
Qwen2.5-32B-Instruct	71.4	61.5	55.9	62.2	62.2
SciJUDGE-Qwen2.5-32B	85.4 (+14.1)	77.9 (+16.4)	82.2 (+26.3)	89.9 (+27.7)	83.7 (+21.4)
Llama3.1-8B-Instruct	34.5	44.3	35.9	35.3	36.8
SciJUDGE-Llama3.1-8B	56.8 (+22.3)	59.8 (+15.6)	55.2 (+19.2)	60.5 (+25.2)	57.3 (+20.5)

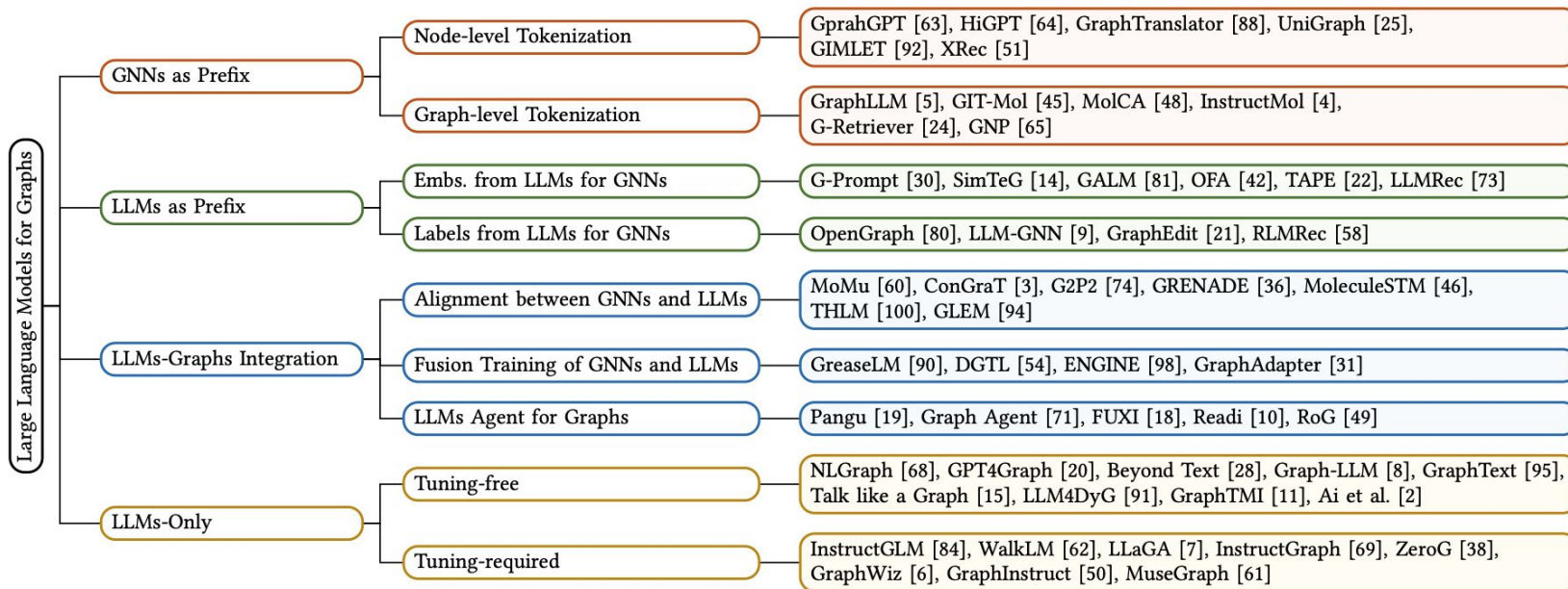
Trending Prediction

- Predict the impact of onsets of ideas.
- Extract 141 features for each pair are calculated.
 - 41 network features
 - 58 of these are node citation features
 - 42 features are about vertex pairs
- Graph Learning?



train dataset: 2016 -- 2019; test dataset: holdout data 2016 -- 2019; eval dataset: 2019 --2022

LLMs + Graph?



Ren et al., [A Survey of Large Language Models for Graphs](#), KDD, 2024.

Takeaway – Graph Based System

- Main Function
 - Relationships between research papers
 - Explore knowledge structures
- Key techniques
 - Citation & Co-Authorship Networks
 - Graph Visualization & Navigation
 - Trending & Citation Analysis
- Challenges
 - Effective Graph Representation
 - Graph Update
 - Integration with LLMs

Agenda

- ❑ Background
- ❑ AI-Enhanced Search System
- ❑ Paper Chat and Scientific QA
- ❑ Graph Based System
- ❑ **Recommendation System**
- ❑ Takeaway

Recommendation System

Papers Communities Organizations Events Blog Pricing

Search

Sign up Login →

Stay Updated on the Trends, Connect with AI Researchers.

#arXiv papers on AI



Upcoming ResearchTrend Connect Sessions

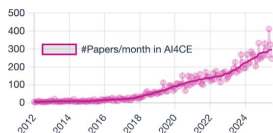
Date	Community	Location	Event
------	-----------	----------	-------

No events available

Feedback

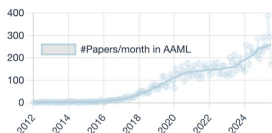
All communities

AI4CE AI for Computational... Follow 3



Papers 14.6K # Contributors 46.6K # followers 3

AAML Adversarial Attacks in... Follow 42



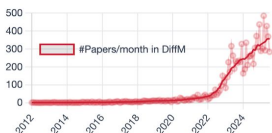
Papers 13.9K # Contributors 31.4K # followers 42

VLM Vision-Language Models Follow 150



Papers 12.0K # Contributors 38.9K # followers 150

DiffM Diffusion Models Follow 116



Papers 10.1K # Contributors 27.9K # followers 116

Recommendation System

The image is a collage of AI-related content. On the left, there's a 'bytez Feed' interface with a search bar, a 'Feed' title, and a main article titled 'AI moves fast Stay faster'. Below this are three filter checkboxes: 'Spot breakthroughs the day they drop', 'See trends before the trend-line', and 'Cut the noise, keep the edge'. A 'Create your feed' button is at the bottom. To the right of the feed are several research paper titles and abstracts: 'AI Flow: Perspectives, Scenarios, and Approaches', 'MiniMax-M1: Scaling Test-Time Compute Efficiently with Lightning Attention', 'Cartridges: Lightweight and general-purpose long context representations via self-study', 'The Lock-in Hypothesis: Stagnation by Algorithm', 'On the Emergence of Linear Analogies in Word Embeddings', and 'Guided Search Strategies in Non-Serializable Environments with Applications to Software Engineering Agents'. At the bottom, there's a paper titled 'MMLU-ProX: A Multilingual Benchmark for Advanced'. On the far right, there are technical diagrams showing 'KV Cache' and 'LLM + KV Cache' with memory usage metrics.

Recommendation System

bytez

notenum.ai

Log in [Sign up](#)

Checking new Arxiv Papers in just 10 minutes

Notenum.ai is the personal newsletter informing you of new research updates. Notenum.ai learns your preferences and curates a short but accurate list of related papers that fit your interests.

[Register for Free](#)

k[1]	k[2]	k[3]	k[4]
v[1]	v[2]	v[3]	v[4]



Recommendation System

The screenshot displays a web application interface for a recommendation system, overlaid on a blurred background of a research paper. The interface is organized into several panels:

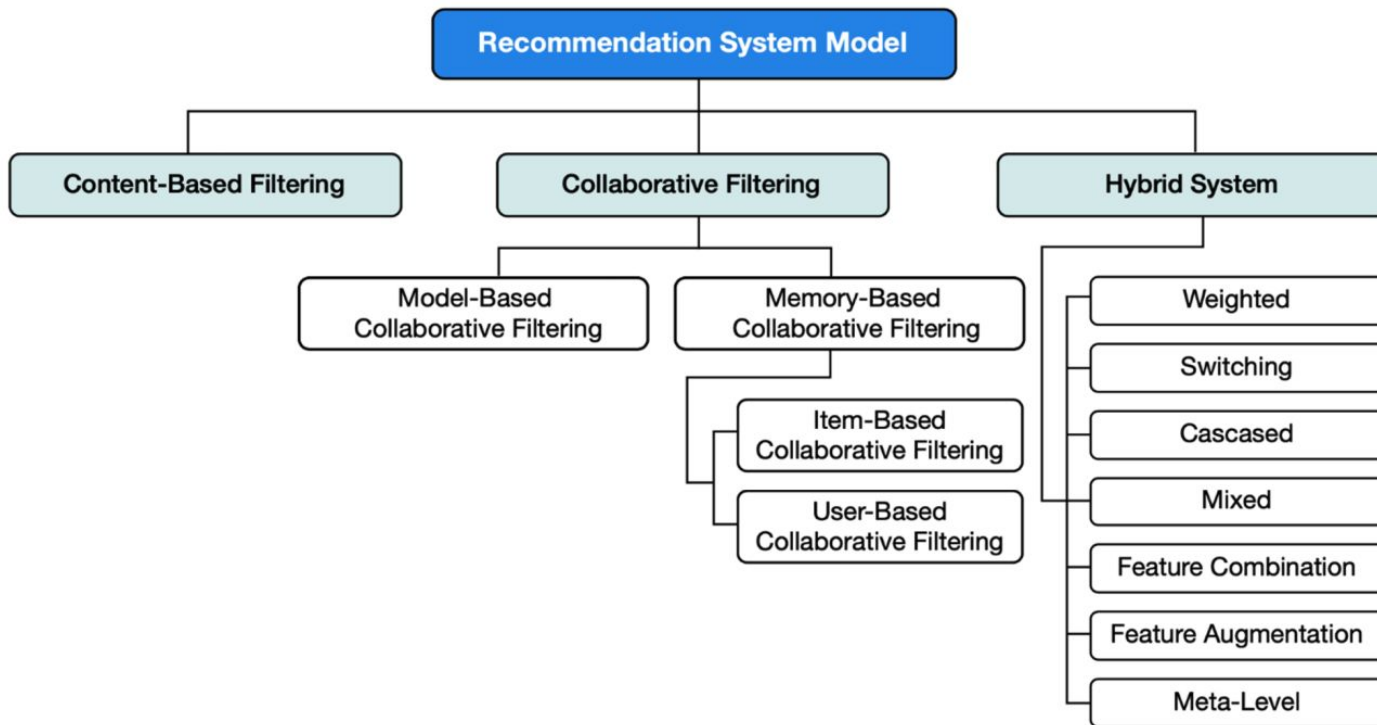
- Left Panel:** A sidebar with navigation options like "Papers", "Communities", "Connections", "Events", "Plans", and "Reviews". It features a "Stay Up" section with a bar chart for "#arXiv papers" and an "All communities" section with a bar chart for "#Papers 14.6K #".
- Top Panel:** A navigation bar with the "bytez" logo, a search bar, and utility buttons for "Log in" and "Sign up".
- 1 selected paper:** A detailed view of a paper by Gonet-Schoupsky and Garip (2016) titled "Laughter and humour interventions for well-being in older adults...". It includes an abstract, objectives, and a list of keywords.
- 1 selected author:** A profile for Guilan Garip, showing 213 citations and a "Collaborators" button.
- Collaborators:** A list of 40 authors, including Freda N. Gonet-Schoupsky, Lucy Yardley, David Sheffield, Robert Bridger, Ainslea Cross, Heather Buchanan, Robert S. Bridger, Miles Richardson, and Aimee Aubeuluck.
- Connections between 40 authors:** A network graph showing the relationships between the listed authors, with Guilan Garip at the center.

Recommendation System

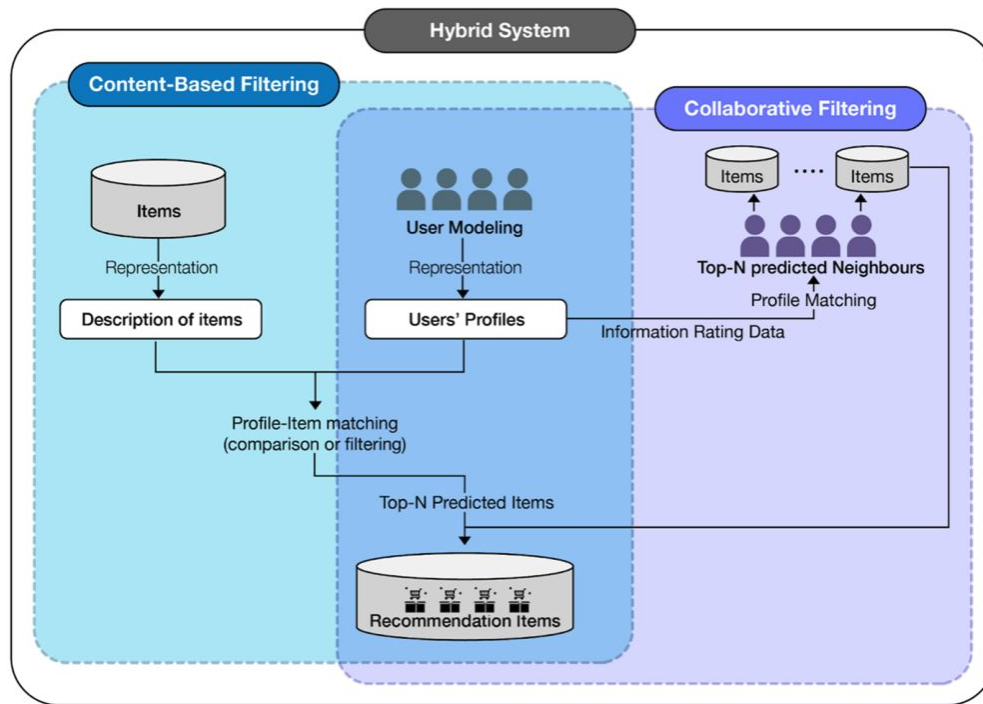
The screenshot displays the notesum.ai web application interface. At the top, there are navigation tabs for Papers, Communication, Conferences, Events, Plans, and Profiles. A search bar is visible with the text "Search". Below the navigation, there are several panels:

- Left Panel:** "Stay Up" section with a bar chart showing "#arXiv papers" from 2012 to 2013. Below it, "All communit" section with a bar chart showing "#Papers 14.6K #".
- Top Right Panel:** "Log in" and "Sign up" buttons.
- Main Content Area:**
 - 1 selected paper:** "Gonot-Schoupsky, Garip, 2016. Laughter and humour interventions for well-being in older adults. A systematic review and intervention classification." The abstract discusses the potential of laughter and humour interventions to increase well-being in a general population of adults aged 65 plus.
 - 1 selected author:** "Guilan Garip, University of Derby, 213 citations." A button says "Click an author to explore their profile".
 - Collaborators:** "40 authors". A button says "Click an author to explore their profile".
 - Connections between 40 authors:** A network visualization area.
- Bottom Panel:** "New conference proceedings" section. It contains two entries:
 - CiteLab: Developing and Diagnosing LLM Citation Generation Workflows via the Human-LLM Interaction** by Jiajun Shen, Tong Zhou, Yubo Chen, Kang Liu, Jun Zhao. Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations). 89 likes, 0 comments, 0 shares. Language.
 - FineCite: A Novel Approach For Fine-Grained Citation Context Analysis** by Lasse M. Jantsch, Dong-Jae Koh, Seonghwan Yoon, Jisu Lee, Anne Lauscher, Young-Kyoon Suh. Findings of the Association for Computational Linguistics: ACL 2025. 83 likes, 0 comments, 0 shares. Language.

Recommendation System

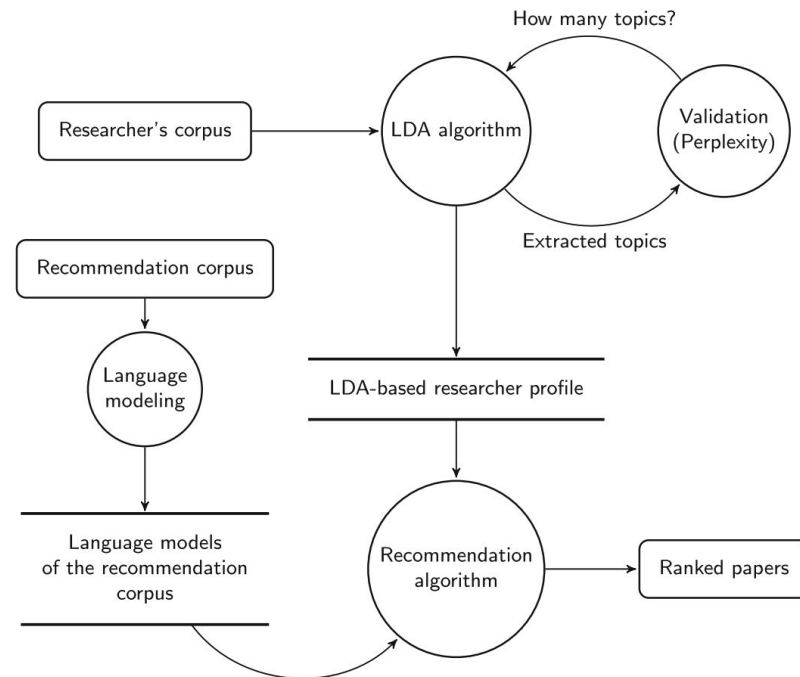


Pipeline of recommendation systems



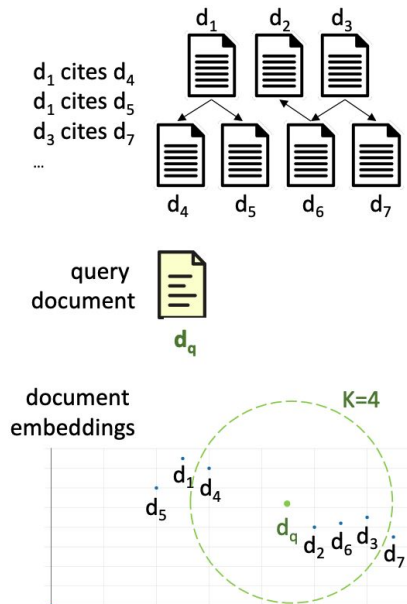
Content-based Methods

- Compares topics from the researcher's profile with the language models of unseen papers.
- Uses the symmetrized Kullback-Leibler divergence to measure similarity between probability distributions (topics and language models).
- **Limitations**
 - Shifted Cold-Start Problem
 - Potential for Limited/General Concepts
 - Lack of Contextual Citation Information



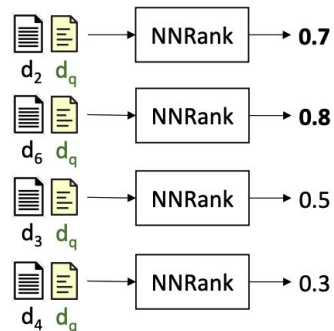
NNSelect+NNRank

Phase 1: candidate selection

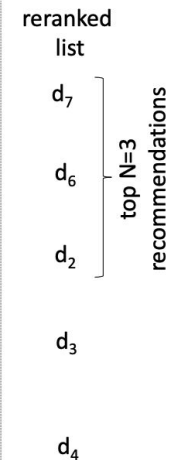
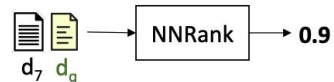


Phase 2: reranking

nearest neighbors of d_q :

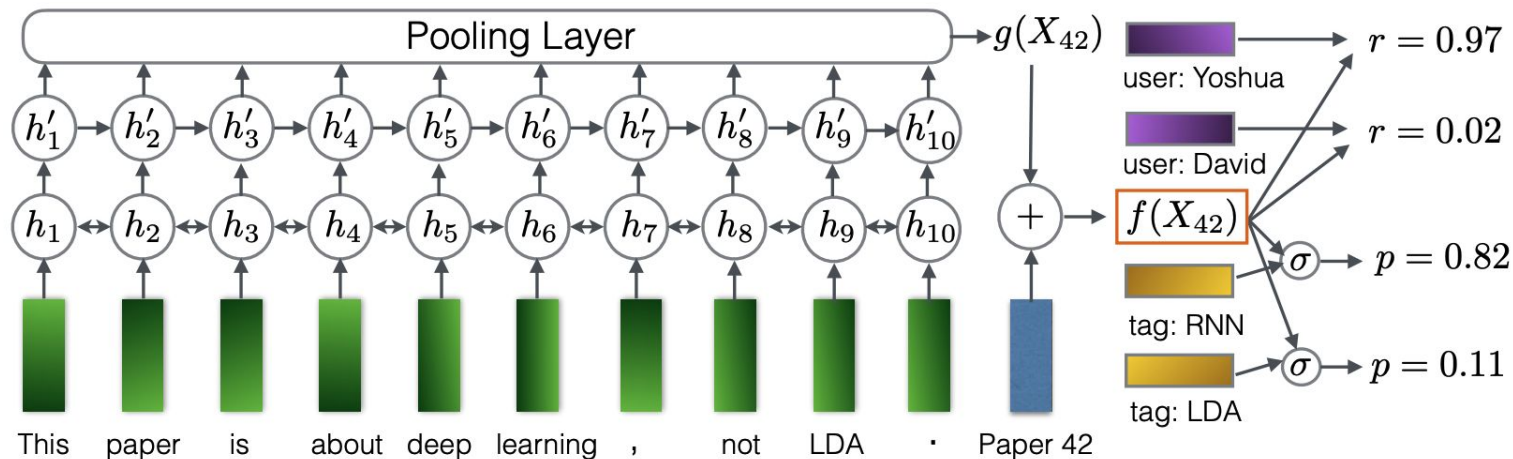


cited in nearest neighbors:

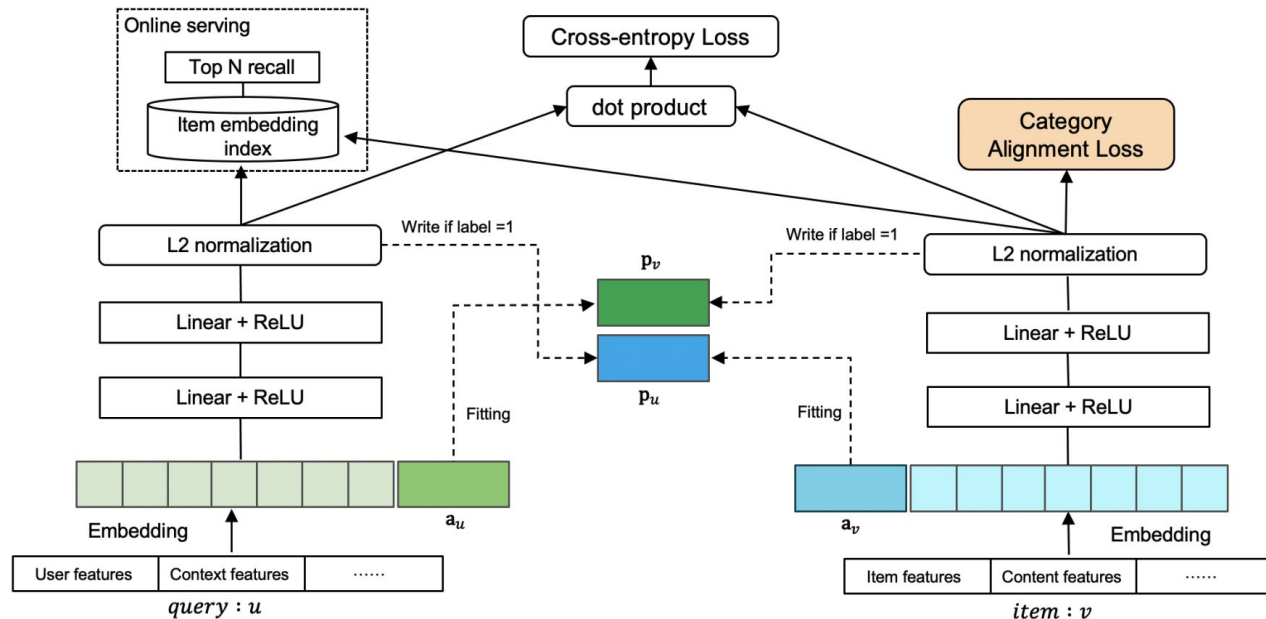


Collaborative Filtering

- By training a text encoder network as a combination of content recommendation and item metadata prediction (e.g., tag prediction)



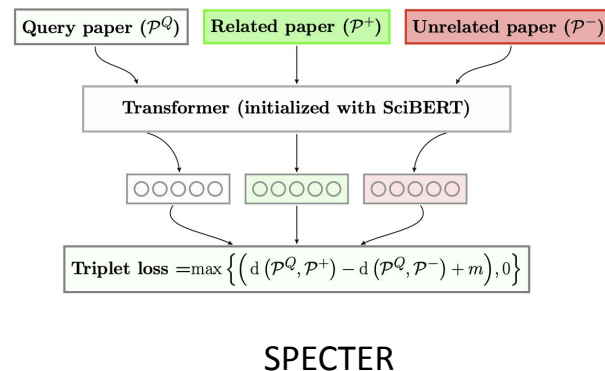
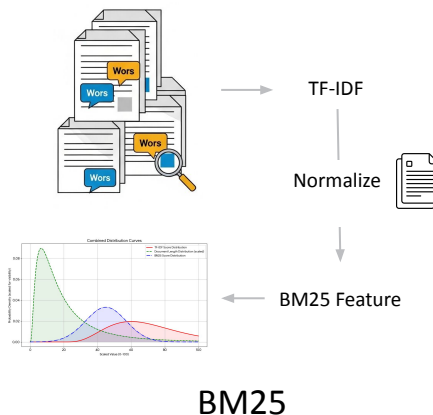
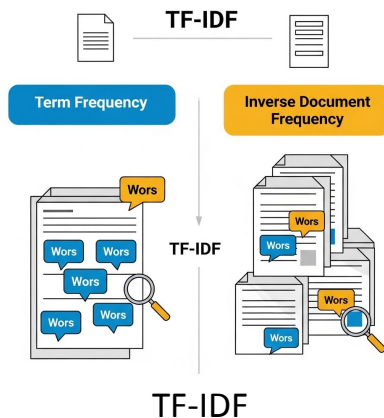
Hybrid Systems



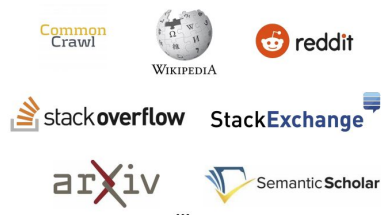
Comparison of Recommendation Systems

Model Type	Pros	Cons
Content-Based Filtering	<ul style="list-style-type: none"> • Works well for new users • Personalized to individual preferences • Doesn't require large user base • Easy to interpret 	<ul style="list-style-type: none"> • Needs detailed item features • Limited discovery of new interests
Collaborative Filtering	<ul style="list-style-type: none"> • Doesn't need item metadata • Leverages collective user behavior 	<ul style="list-style-type: none"> • Struggles with sparse data • Scalability issues in memory-based
Hybrid Systems	<ul style="list-style-type: none"> • Combines best of all models • Mitigates individual weaknesses • Often improves accuracy 	<ul style="list-style-type: none"> • Increased complexity • Requires more computation

Document Embeddings

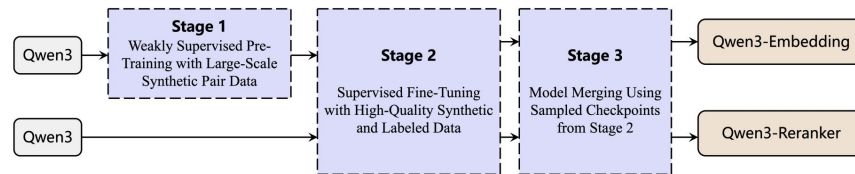
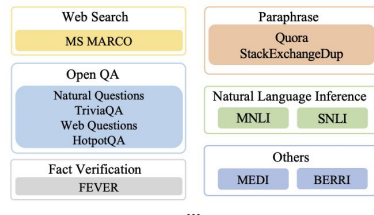


Unsupervised Contrastive Pre-training on Massive Text Pairs mined from the Web



GTE

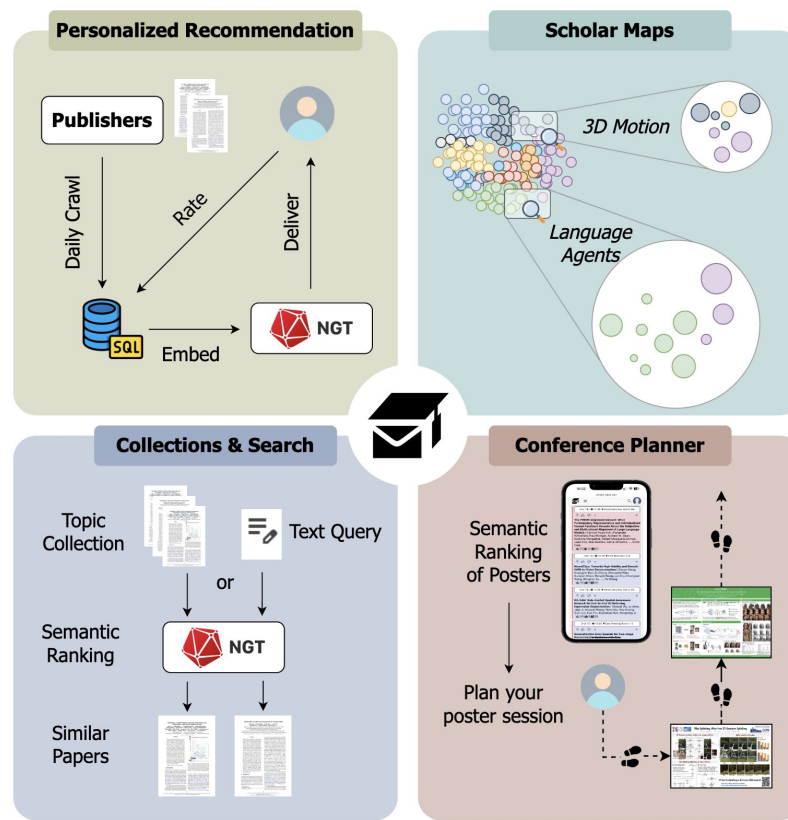
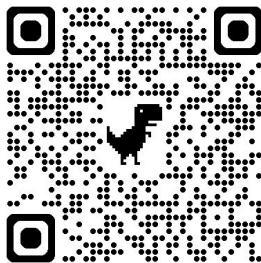
Supervised Contrastive Fine-tuning on Annotated Text Triples from Multiple Tasks



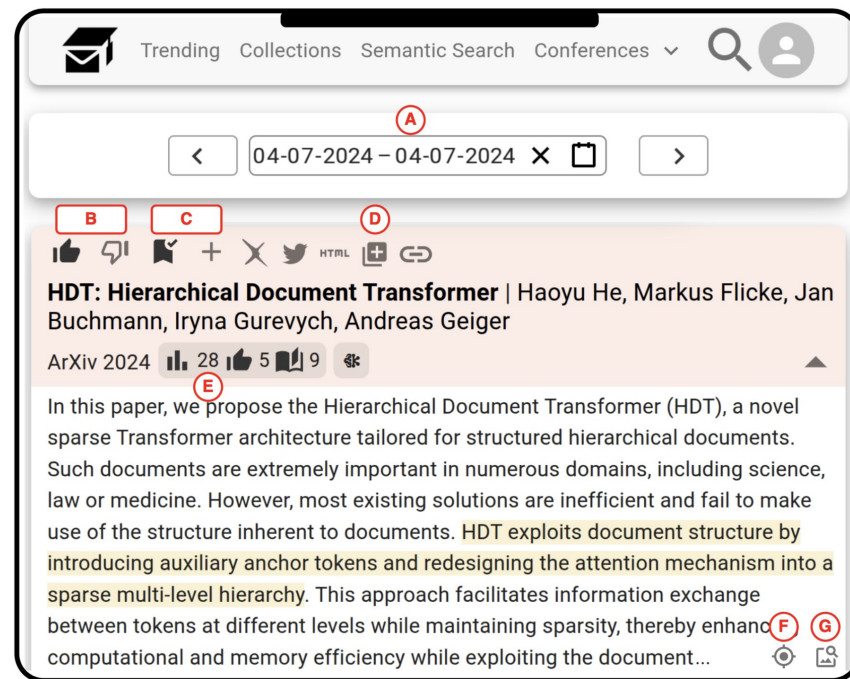
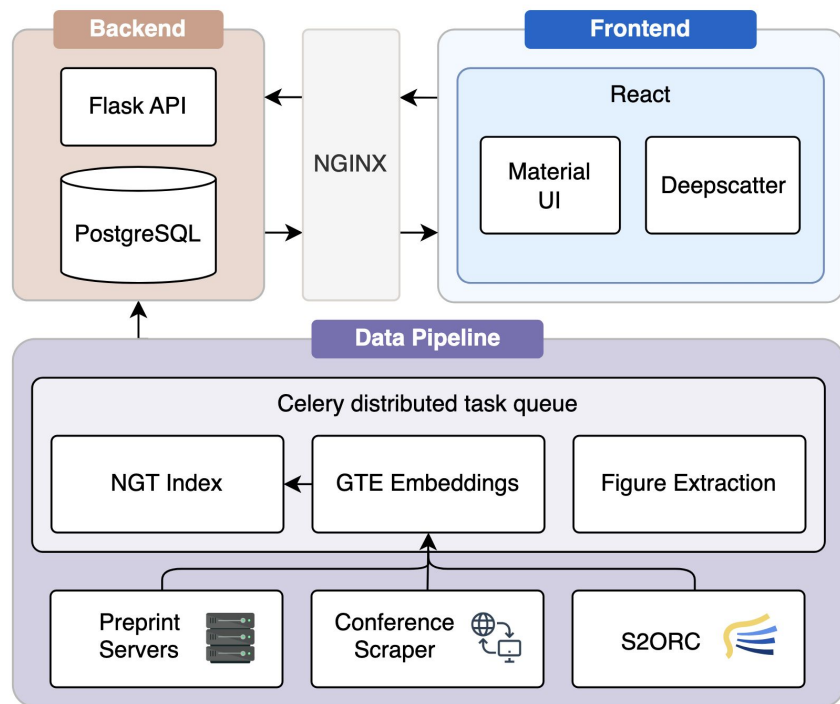
Qwen-3 Embedding

Scholar Inbox

- Personalized Recommendations
- Scholar Maps
- Collections
- Conference Planner



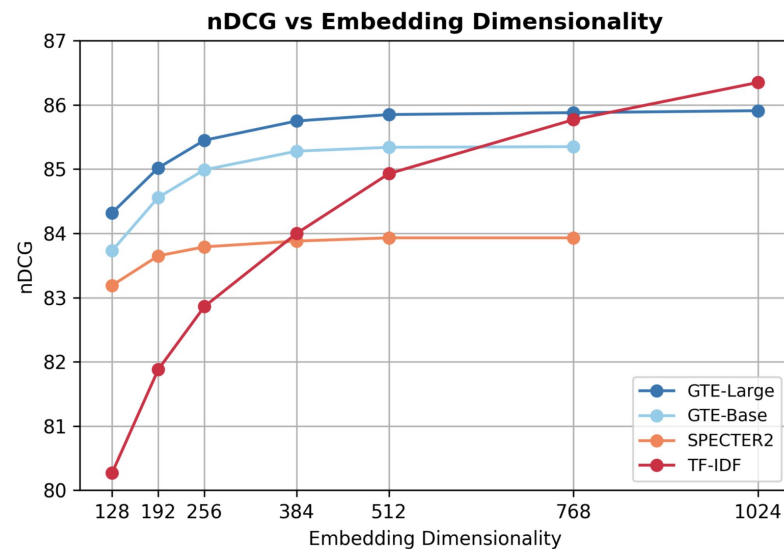
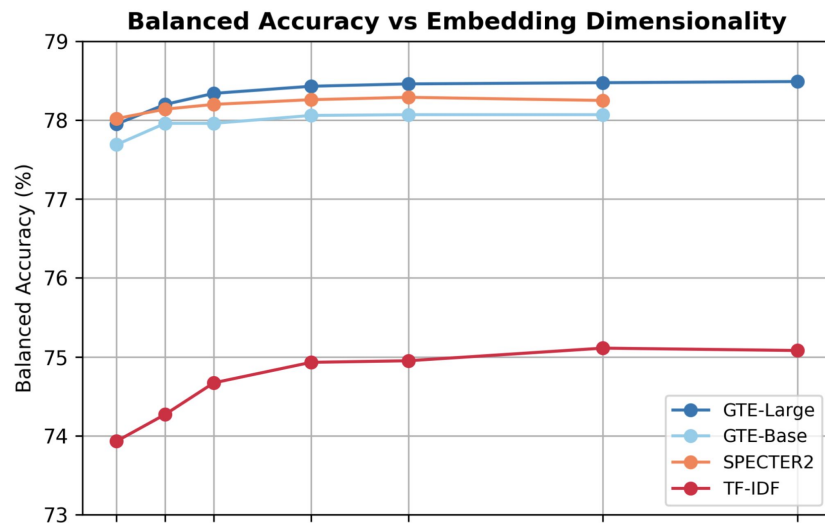
Software Flow



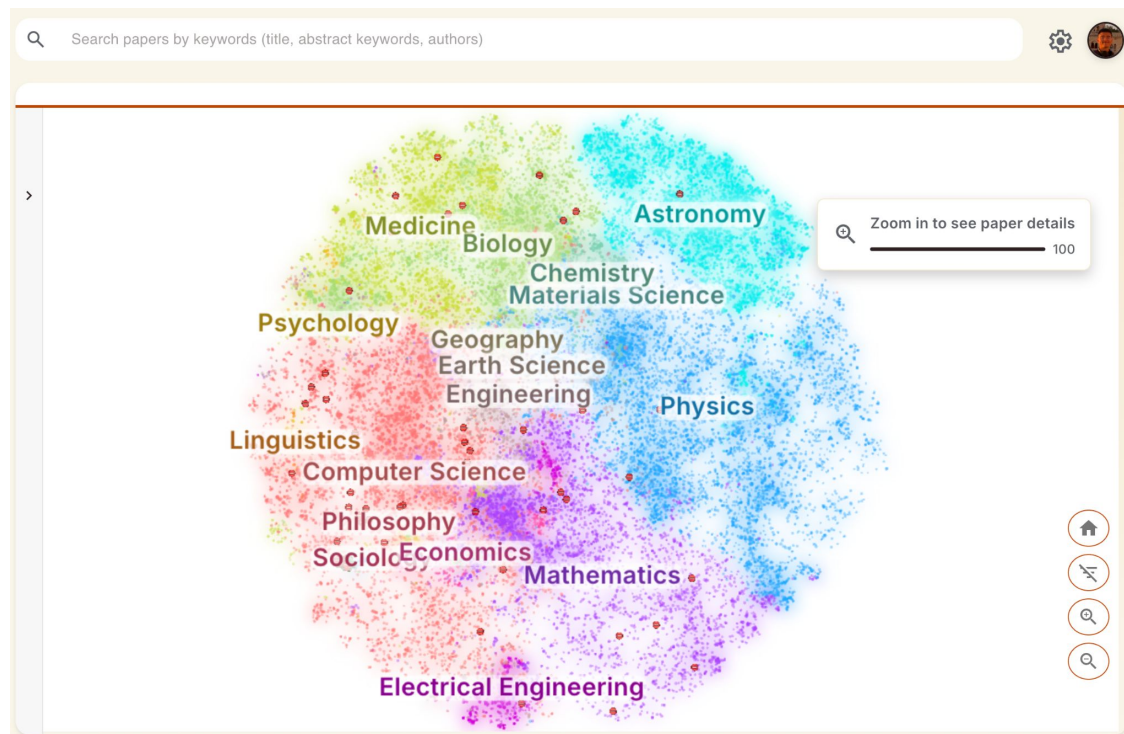
Recommender System

The screenshot displays the Scholar Inbox interface. On the left, there is a navigation sidebar with options like 'Digest', 'Trending', 'Interactions', 'Scholar Maps', 'Conferences', and 'Collections'. The main content area features a search bar at the top, a home button, a date indicator for '19 Mar 2026', and a paper count of '5/1434 Papers'. The featured paper is 'Public Profile Matters: A Scalable Integrated Approach to Recommend Citations in the Wild' by Karan Goyal, Dikshant Kukreja, and Vikram Goyal. It includes a thumbnail of the paper, a title, authors, submission date, and a list of tags such as 'Computer Science', 'Information Retrieval', and 'Citation Recommendation'. The abstract text is visible, highlighting the paper's contribution to citation recommendation systems. On the right side, there are interactive options like 'Bookmark', 'Details', 'Collections', 'arXiv', 'PDF', 'HTML', and 'Share'. At the bottom, there are navigation icons for 'AI Summary', 'Figures & Tables', 'Scholar Maps', and 'Similar Papers'.

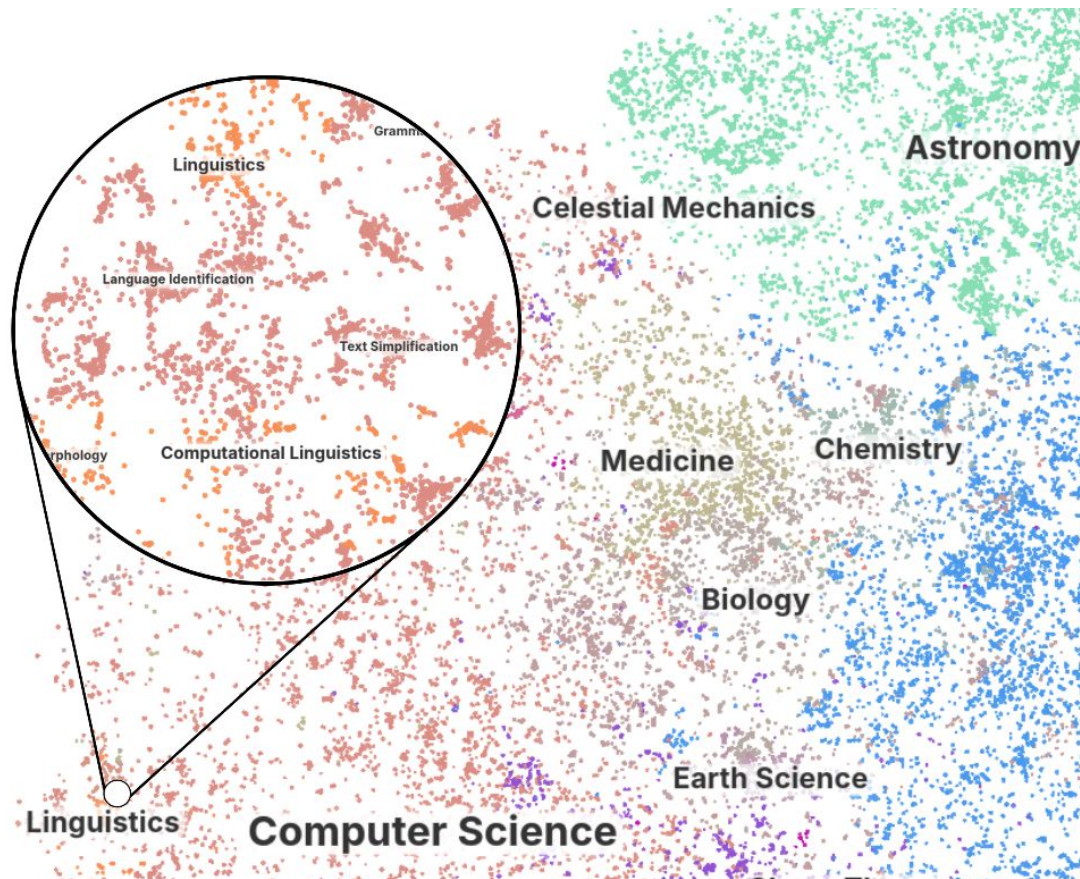
Embeddings



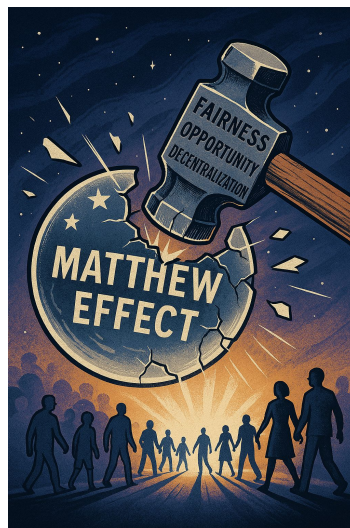
Scholar Maps



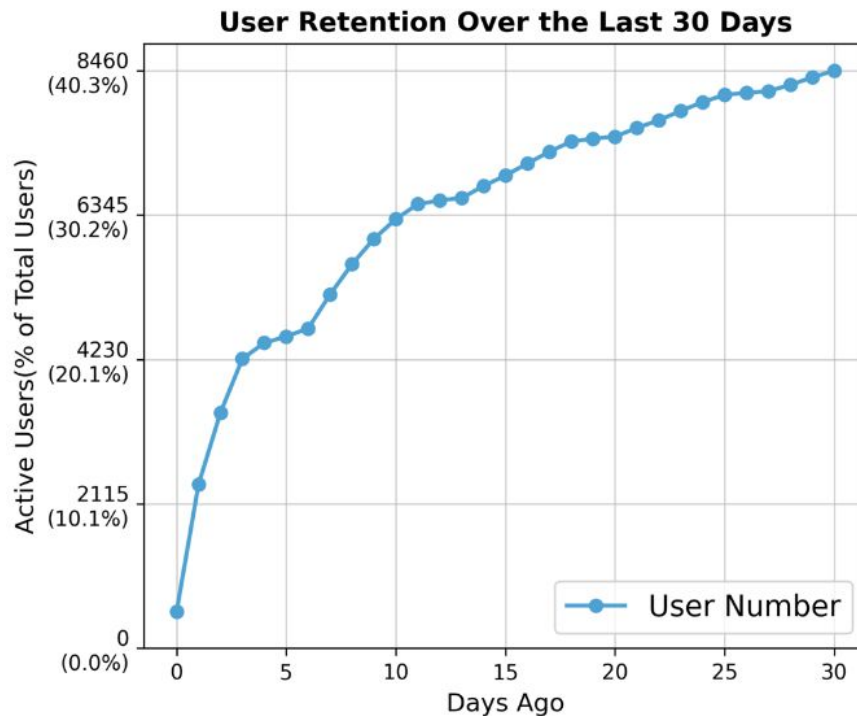
Scholar Maps



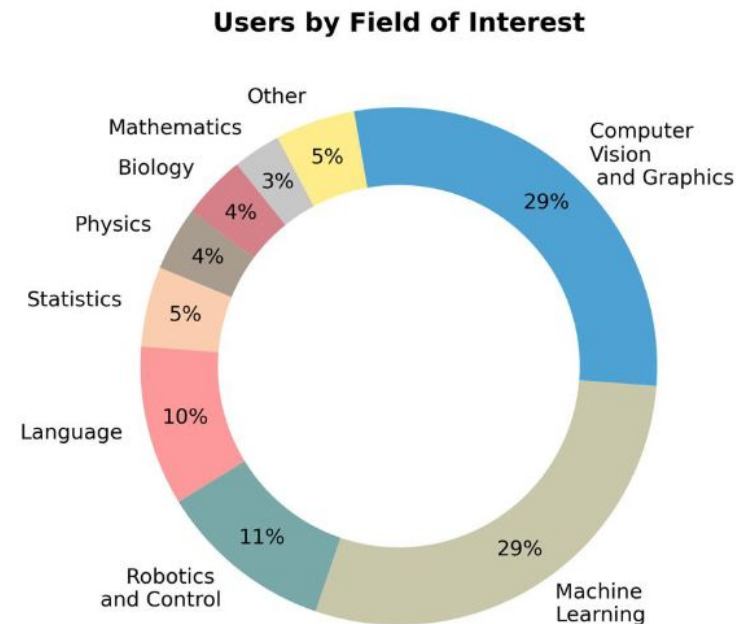
Why we create Scholar Inbox?



9,000 Users Come Back Again and Again

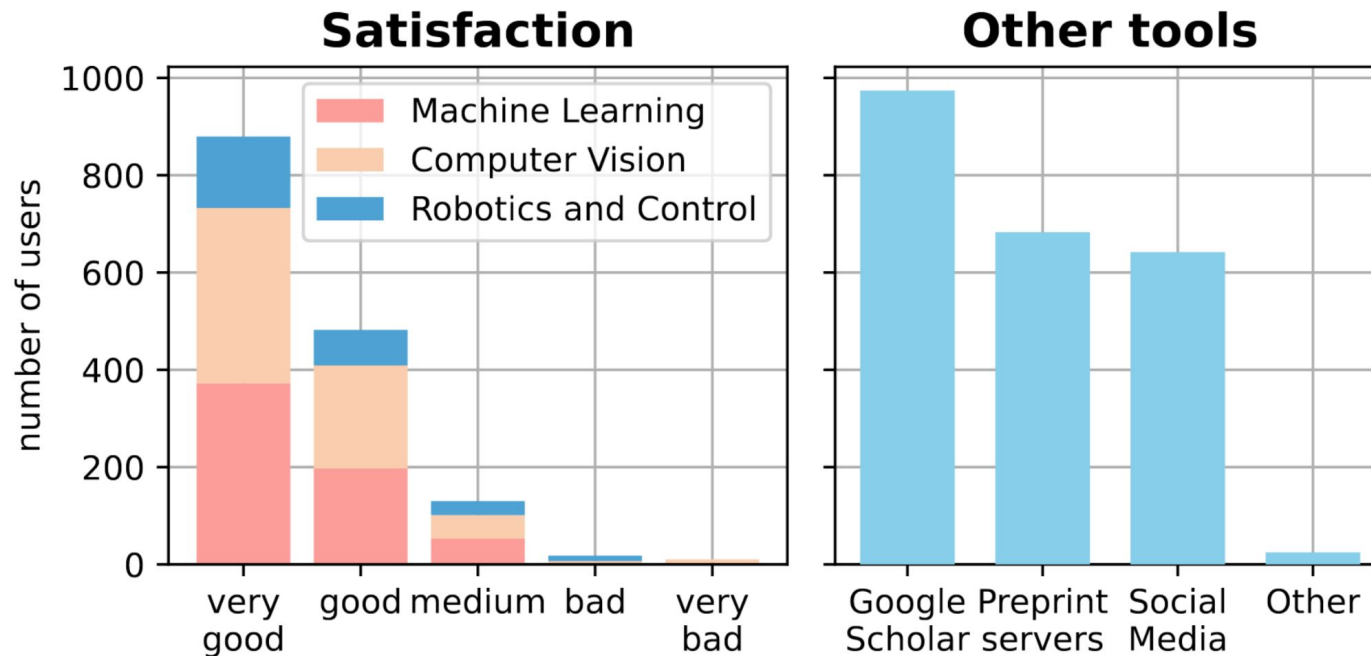


(a) User Retention



(b) User Domain Distribution

User Study



Takeaway – Recommendation System

- Main Function
 - Keep updated literatures
 - Personalized Recommendation
 - Collections
- Key techniques
 - Content-Based Filtering
 - Collaborative Filtering
 - Semantic Search & Embeddings
- Challenges
 - Cold Start Problem
 - Overpersonalization vs Matthew Effect
 - Dynamic Interests of Researchers

Overview of popular literature search

	Platform	Search	Recommendations	Citation Analysis	Trending Analysis	Author Profiles	Visualization Tools	Paper Chat	Idea Generation	Paper Writing	Summarization	Paper Review	Datasets	Code Repositories	LLM Integration	Web API	Personalization	Cost	Data Source
AI-Enhanced Search	Elicit	✓					✓	✓						✓				Freemium	125 million
	OpenScholar	✓	✓				✓	✓						✓				Free	45 million
	Undermind	✓	✓				✓	✓						✓	✓			Premium	over 200 million
	Perplexity	✓					✓	✓										Freemium	
	Consensus	✓	✓				✓	✓						✓	✓			Freemium	over 200 million
	SciSpace	✓	✓				✓	✓						✓				Freemium	
	scienceQA	✓	✓	✓			✓	✓						✓				Freemium	220 million
	PaperQA2	✓					✓	✓						✓	✓			Free	
	Paperguide	✓	✓				✓	✓							✓			Freemium	
	HyperWrite	✓					✓	✓	✓	✓	✓				✓			Premium	
ResearchKick	✓					✓	✓	✓	✓	✓				✓	✓		Premium		
Graph-Based	Connected Papers	✓	✓			✓												Freemium	214 million
	ScholarGPS	✓		✓	✓	✓												Free	over 200 million
	CiteSpace			✓	✓	✓												Freemium	
	Sci2				✓	✓												Free	
	NLP KG	✓	✓	✓	✓	✓												Free	
ORKG ASK	✓	✓						✓						✓			Free	76 million	
Paper Chat	ChatGPT	✓					✓	✓	✓	✓	✓			✓	✓			Freemium	10 pdf files
	Claude	✓					✓	✓	✓	✓	✓			✓	✓			Freemium	5 pdf files
	Deepseek	✓					✓	✓	✓	✓	✓			✓	✓			Free	
	Research		✓				✓	✓	✓	✓	✓			✓				Freemium	1 pdf file
	NotebookLM						✓	✓	✓	✓	✓			✓		✓		Freemium	50 pdf files
	Enago Read	✓	✓				✓	✓	✓	✓	✓			✓				Freemium	1 pdf file
	DocAnalyzer.AI		✓				✓	✓	✓	✓	✓			✓	✓	✓		Premium	few pdf files
	CoralAI		✓				✓	✓	✓	✓	✓			✓				Freemium	1 pdf file
	ExplainPaper						✓	✓	✓	✓	✓			✓				Freemium	1 pdf file
ChatPDF	✓	✓				✓	✓	✓	✓	✓			✓				Premium	1 pdf file	
Recommender	Arxiv Sanity	✓	✓	✓														Free	
	Scholar Inbox	✓	✓	✓										✓	✓			Free	
	ResearchTrend.ai	✓			✓	✓												Freemium	
	TrendingPapers	✓	✓		✓									✓	✓			Free	
	Bytez	✓			✓				✓	✓				✓	✓			Freemium	
	Notesum.ai	✓	✓	✓							✓			✓	✓			Freemium	
	Research Rabbit	✓	✓		✓													Free	

Overview of popular literature search

	Platform	Search	Recommendations	Collections	Citation Analysis	Trending Analysis	Author Profiles	Visualization Tools	Paper Chat	Idea Generation	Paper Writing	Summarization	Paper Review	Datasets	Code Repositories	LLM Integration	Web API	Personalization	Cost	Data Source
Search Engines	Google Scholar	✓	✓	✓	✓	✓									✓				Free	
	Semantic Scholar	✓	✓	✓	✓	✓	✓	✓		✓					✓	✓	✓		Free	214 million
	Baidu Scholar	✓	✓	✓	✓	✓	✓								✓	✓			Freemium	680 million
	BASE	✓		✓												✓			Free	415 million
	Internet Archive Scholar	✓														✓			Free	35 million
	Scilit	✓		✓			✓												Free	172 million
	The Lens	✓		✓				✓									✓		Freemium	284 million
	Science.gov	✓																	Free	several million
	Academia.eu	✓		✓				✓											Freemium	55 million
	OpenAlex	✓						✓										✓	Freemium	
	AceMap	✓			✓	✓	✓	✓						✓					Free	260 million
	PubTator3	✓		✓	✓												✓		Free	6 million
	Benchm.	Papers with Code	✓									✓	✓							Free
ScienceAgentBench										✓		✓	✓	✓					Free	
ORKG Benchmarks					✓		✓					✓							Free	
Huggingface		✓		✓			✓					✓	✓						Freemium	

Agenda

- ❑ Background
- ❑ AI-Enhanced Search System
- ❑ Paper Chat and Scientific QA
- ❑ Graph Based System
- ❑ Recommendation System
- ❑ **Takeaway**

Takeaway

- AI is transforming literature search.

Takeaway

- AI is transforming literature search.
- AI tools boost discovery but still require oversight.

Takeaway

- AI is transforming literature search.
- AI tools boost discovery but still require oversight.
- Four AI paradigms jointly redefine research workflows.

Takeaway

- AI is transforming literature search.
- AI tools boost discovery but still require oversight.
- Four AI paradigms jointly redefine research workflows.
- Future directions point to smarter, multimodal systems.

Future Direction

- Multimodal literature search
 - Integrating text and figures/tables
- Event-oriented summarization
 - Extracting and organizing key research events (e.g., discoveries, methods, results) for clearer insights
- Real-time updates & knowledge tracking
 - Continuous integration of new findings
- Integration with scientific knowledge graphs
 - Structured, interconnected research data

References

1. Gemini, Gemini Deep Research, 2026.
2. Flicke et al., Scholar Inbox: Personalized Paper Recommendations for Scientists, ACL Demo, 2025.
3. Gu et al., Forecasting high-impact research topics via machine learning on evolving knowledge graphs, Machine Learning: Science and Technology, 2025.
4. Agarwal et al., LitLLMs, LLMs for Literature Review: Are we there yet? Transactions on Machine Learning Research, 2025.
5. Khalid et al., Comprehensive review of academic search systems: evolution, analysis, and future research directions, Social Network Analysis and Mining, 2025.
6. He et al., PaSa: An LLM Agent for Comprehensive Academic Paper Search, ACL, 2025.
7. OpenAI, Deep Research System Card, 2025.
8. Li et al., ChemVLM: Exploring the Power of Multimodal Large Language Models in Chemistry Area, AAAI, 2025.
9. Li et al., BLADE: Enhancing Black-box Large Language Models with Small Domain-Specific Models, AAAI, 2025.
10. Zhang et al., SCITAT: A Question Answering Benchmark for Scientific Tables and Text Covering Diverse Reasoning Types, ACL-findings, 2025.
11. He et al., GIVE: Structured Reasoning of Large Language Models with Knowledge-Graph-Inspired Veracity Extrapolation, ICML, 2025.
12. Giglou et al., LLMs4Synthesis: Leveraging Large Language Models for Scientific Synthesis, JCSL, 2025.
13. Mario et al., Forecasting the future of artificial intelligence with machine learning-based link prediction in an exponentially growing knowledge network, Nature machine intelligence, 2023.
14. Kang et al., Taxonomy-guided Semantic Indexing for Academic Paper Search, EMNLP, 2024.
15. Ren et al., A Survey of Large Language Models for Graphs, KDD, 2024.
16. L´ala et al., PaperQA: Retrieval-Augmented Generative Agent for Scientific Research, arXiv, 2023.
17. Auer et al., The SciQA Scientific Question Answering Benchmark for Scholarly Knowledge, Scientific Reports, 2023.
18. Ko et al., A Survey of Recommendation Systems: Recommendation Models, Techniques, and Application Fields, electronics, 2022.
19. Bhagavatula et al., Content-Based Citation Recommendation, NAACL, 2018.