Google

Academic Paper Search Engine

AI Mode

Group 26

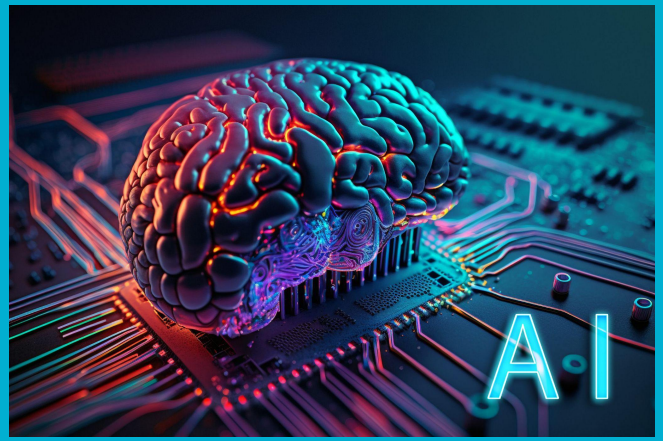
Gordon Yang
Aidan Manternach
Jia-Ming Lin
Cheng-Lun Tai
Tanish Hemadri Talapaneni

Motivation

Motivation: Fast and effective data search is crucial (rise of with RAG for LLMs)

Objective: Create a high quality, state of the art search engine for AI papers

Google Scholar



arXiv Dataset

Total Papers: 2,895,350

AI Papers: 150,863

id	0704.0304
authors	Carlos Gershenson
title	The World as Evolving Information
update_date	2013-04-05
categories	cs.IT cs.AI math.IT q-bio.PE
abstract	This paper discusses the benefits of describing the world as information, especially in the study of the evolution of life and cognition. Traditional studies encounter problems because it is difficult to describe life and cognition in terms of matter and energy, since their laws are valid only at the physical scale. However, if matter and energy, as well as life and cognition, are described in terms of information, evolution can be described consistently as information becoming more complex. The paper presents eight tentative laws of information, valid at multiple scales, which are generalizations of Darwinian, cybernetic, thermodynamic, psychological, philosophical, and complexity principles. These are further used to discuss the notions of life, cognition and their evolution.
doi	10.1007/978-3-642-18003-3_10

Data Preparation

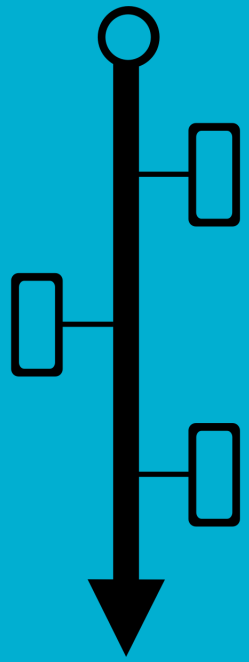
- Categorization: Filtering for cs.AI tags
- Text Cleaning:
  - Removing Punctuation
  - Lower case words
  - Stemming (running, runs, ran -> run)
  - Common Word Removal (the, and, etc.)

Methodology - Levels Of Search Engine

Level 1: TF-IDF

Level 2: Vector Embeddings

Level 3: Deep Learning



Google!

Google

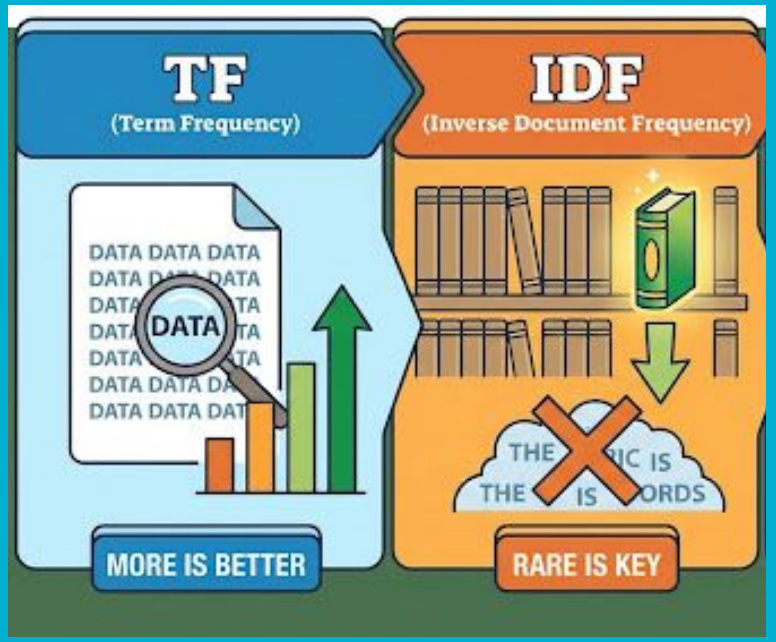
Gemini

Level 1
TF-IDF Search
[Term Frequency – Inverse Document Frequency]

(Classic Statistical)

TF-IDF Search

1. **Rewards Frequency**: Words appearing often in a document get higher scores
2. **Penalises Noise**: Common words (like "the") are downgraded so they don't dominate



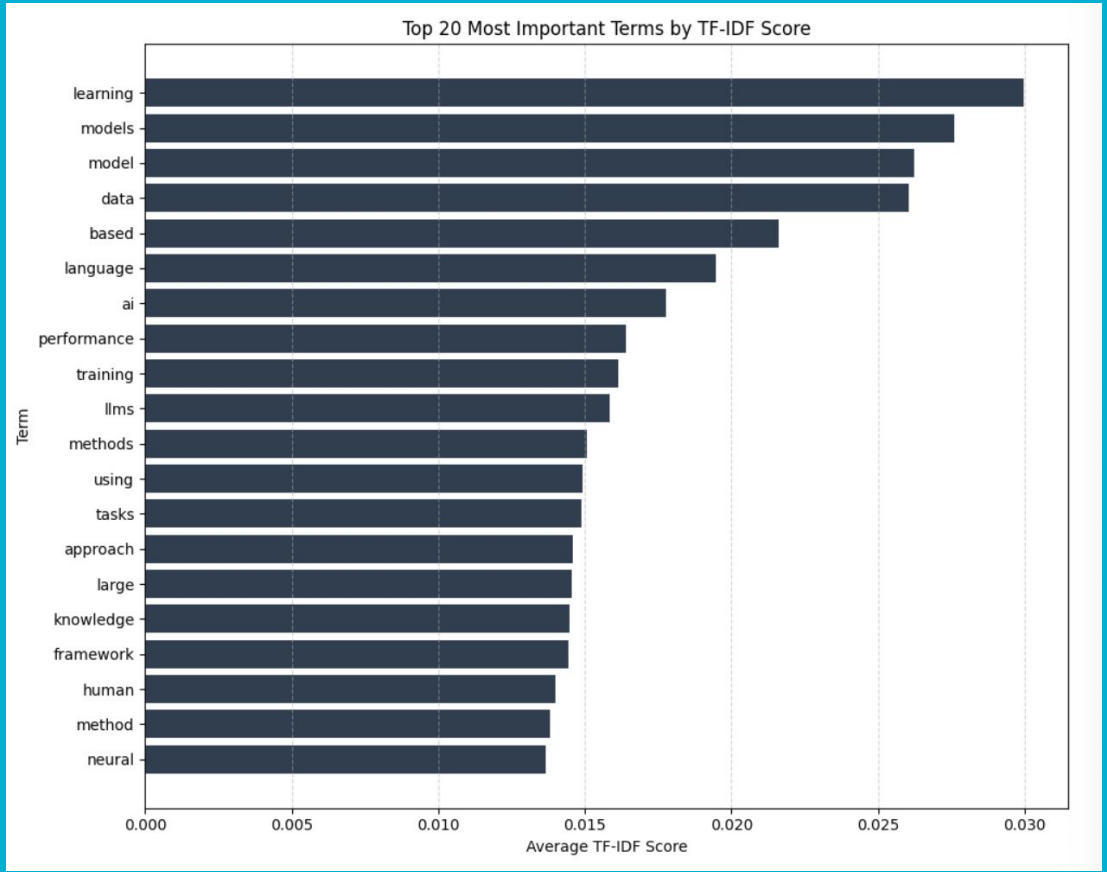
Most Important Terms by TF-IDF Score

Purpose:

- To characterize the content of the research paper dataset
- Identify the most discriminative, high-value keywords used for retrieval.

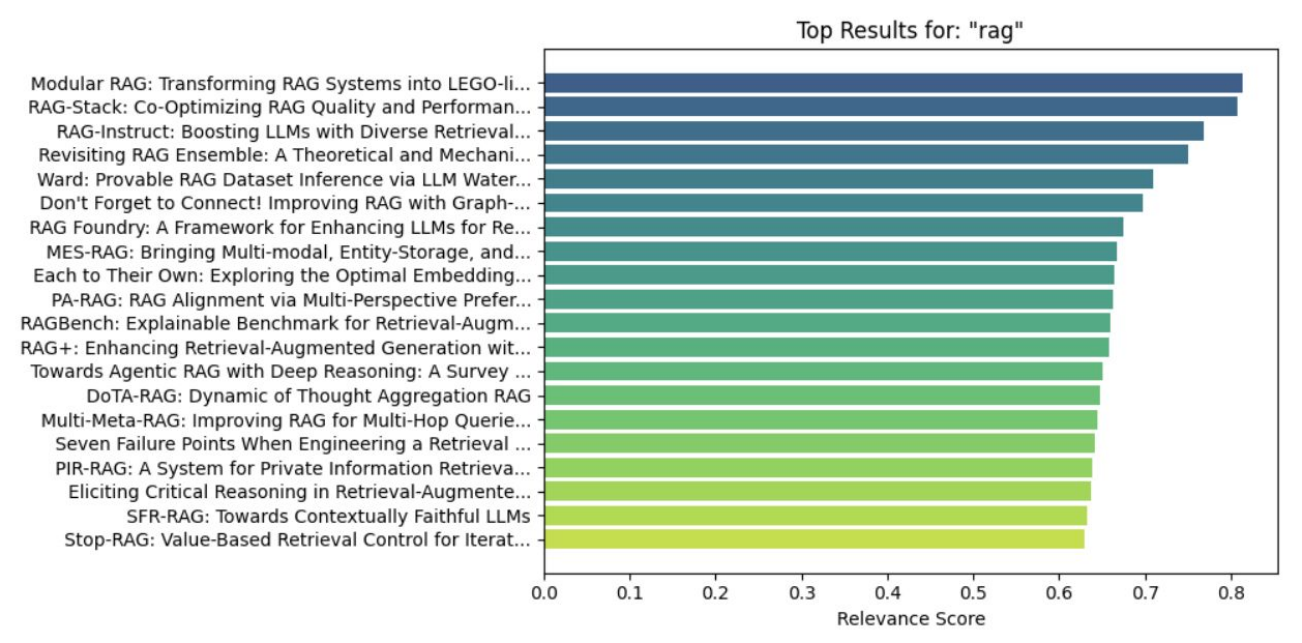
Values:

- Better Search Terms
- Quick Visualize Summary of the Core Topics



Top Results for query: “rag”

Y-axis: Shows the top 20 documents that the TF-IDF retrieval system identified as most relevant to the query “rag”.
X-axis: Shows the cosine similarity value calculated between the query vector and document vector.

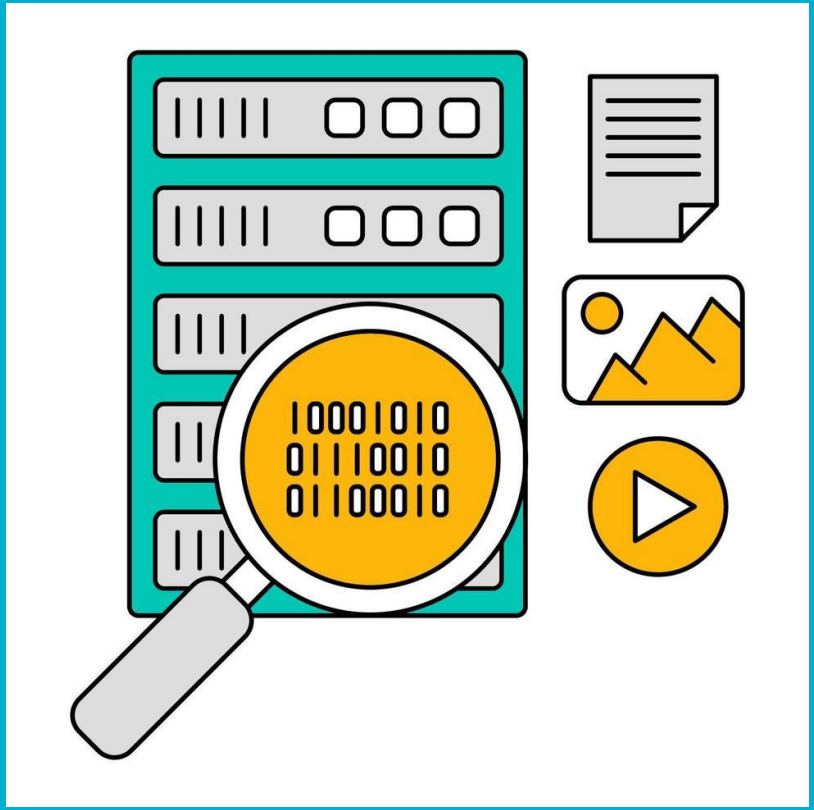


Level 2
BM25 and Vector Search

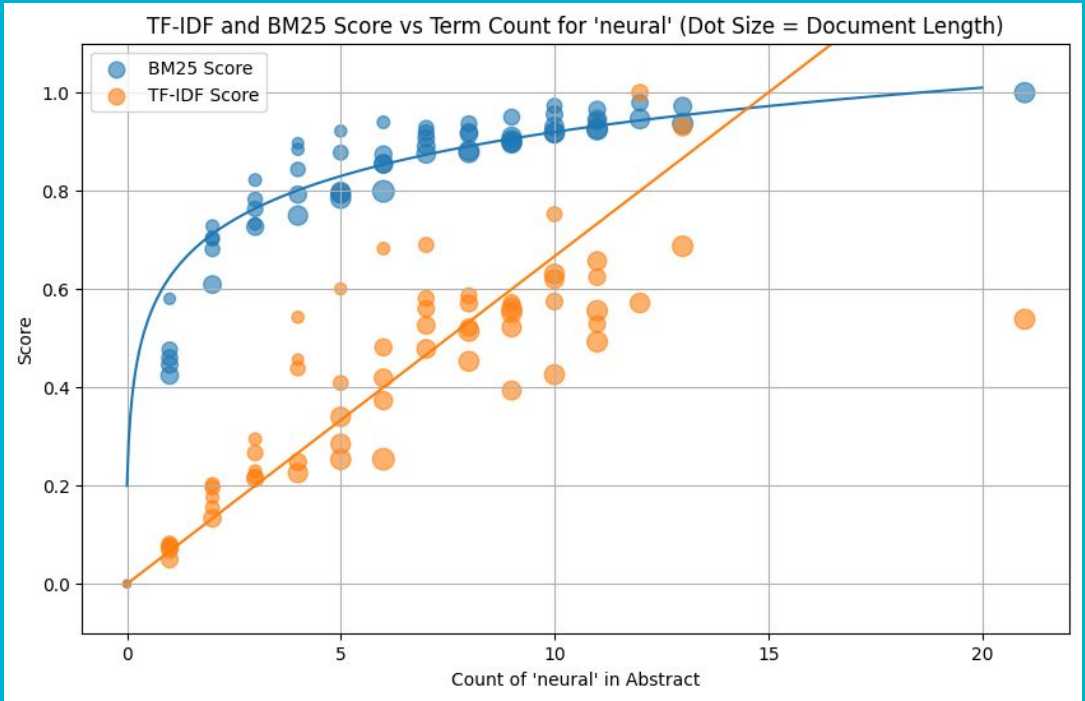
(Semantic and Keyword Search)

BM25 and Vector Search

1. **BM25:** advanced keyword based ranking algorithm
2. **Vector Search:** ranks documents based on embedding distance from query

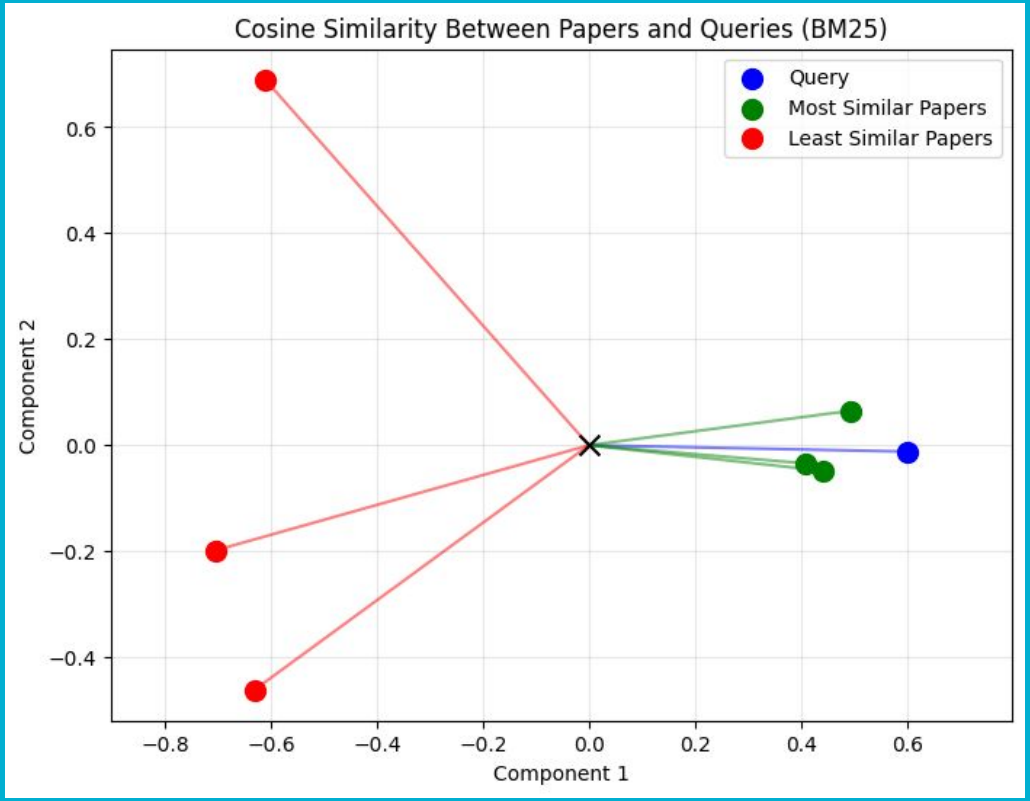


BM25 vs TF-IDF



- BM25 does not linearly increase score with term frequency (term frequency saturation)
- BM25 penalizes long documents to prevent length bias in rankings

Vector Embeddings



Query: Neural Network

BM25 Search

- Matches on keywords but not necessarily semantic meaning

Defending Against Backdoor Attack on Graph Nerual Network by Explainability
2022-09-08
Exploiting Meta-Learning-based Poisoning Attacks for Graph Link Prediction
2025-10-21
Network Engineering for Complex Belief Networks
2013-02-18
Deep Neural Networks as Complex Networks
2022-09-14
Universal Network Representation for Heterogeneous Information Networks
2018-11-30

Vector Embedding Search

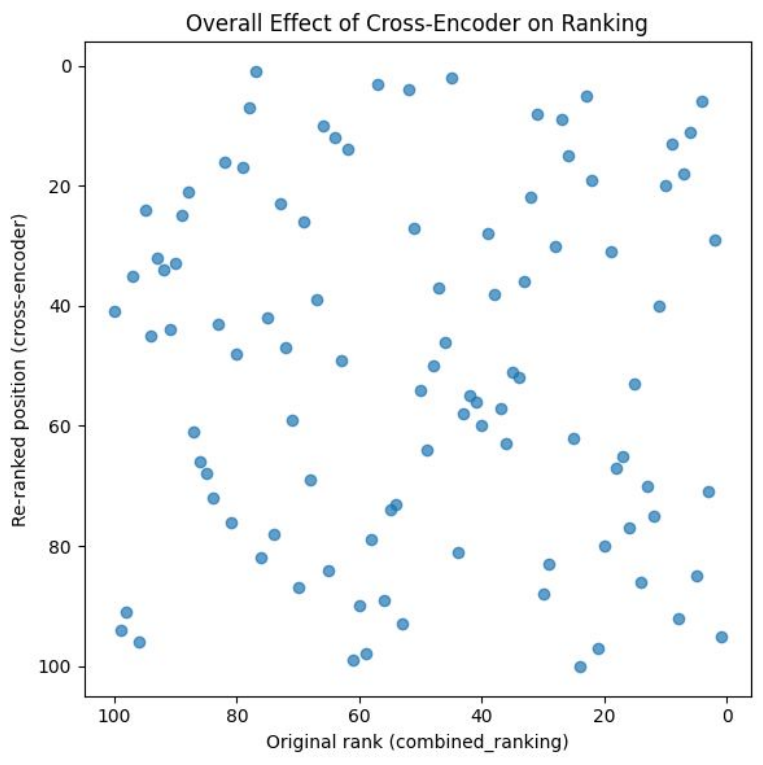
- Models semantic meaning in search

Towards Repairing Neural Networks Correctly
2021-05-07
Interpreting Neural Networks through Mahalanobis Distance
2024-10-28
Assessing Intelligence in Artificial Neural Networks
2020-06-05
Neural Logic Networks
2019-10-22
Architecture Agnostic Neural Networks
2020-12-14

Level 3
Neural Network Reranking

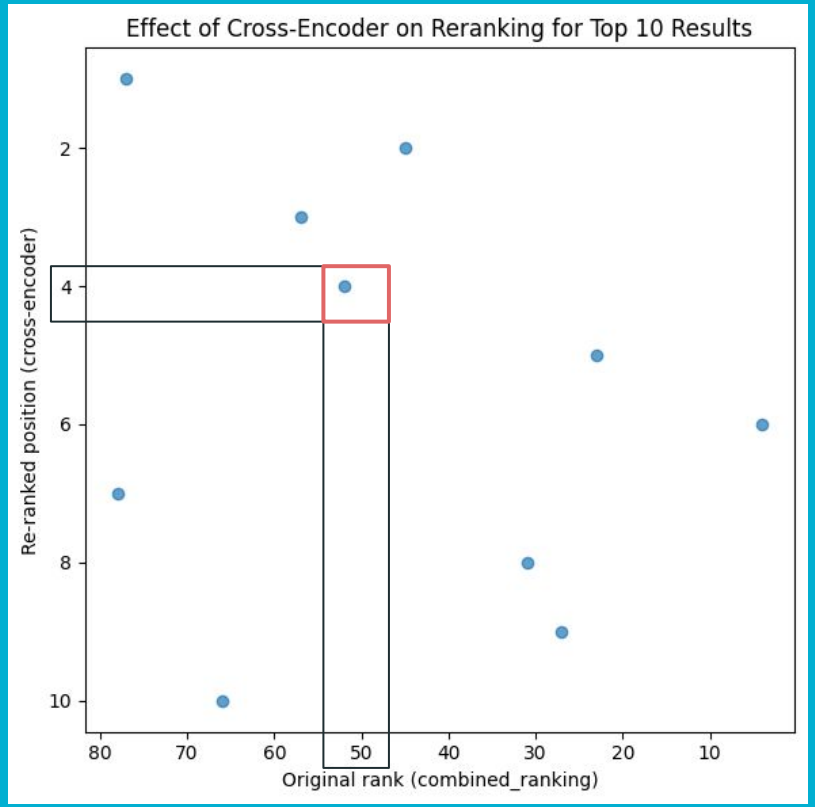
(Using Cross-Encoder)

Reranking Effect of the Cross-Encoder

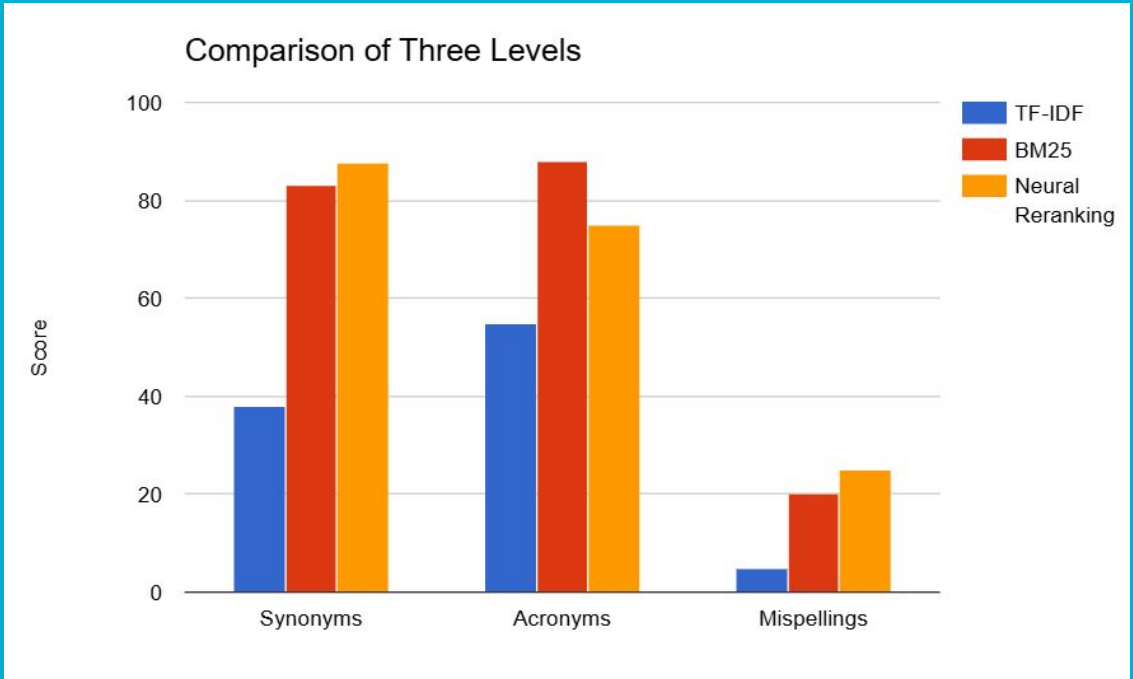


- Using the “ms-marco-MiniLM-L6-V2” cross-encoder from hugging face.
- Promoting papers that were originally ranked low by BM25, but semantically relevant to given query.
- Demonstrated finer-grained relevance judgment of neural reranking than BM25 and dense embeddings

Top 10 Results of Neural Reranking



Search Algorithm Comparison



Queries

Synonym: “Deep Learning for Images”

Acronym: “RAG for Question Answering”

Misspelling: “Retrival Augmnted Generation Models”

Thank You!

Appendix

Synonym Search Comparison

 # Synonyms / Paraphrased Query : deep learning for images

- TF-IDF Search : 38/100
- BM25 and Vector Search : 83/100
- Neural Network Reranking : 87.6/100

Conclusion : As the retrieval level increases, the system shifts from generic deep learning papers (L1) to image-focused deep learning papers (L2), and finally to the most task-relevant image-centric deep learning papers (L3)

Acronym Search Comparison

 # Acronyms vs. Full Terms: RAG for question answering

- TF-IDF Search : 55/100
- BM25 and Vector Search : 88/100
- Neural Network Reranking : 75/100

Conclusion : TF-IDF cannot link acronyms with their full forms, dense retrieval unifies them in embedding space, and neural reranking preserves relevance but depends heavily on the quality of BM25 candidate retrieval.

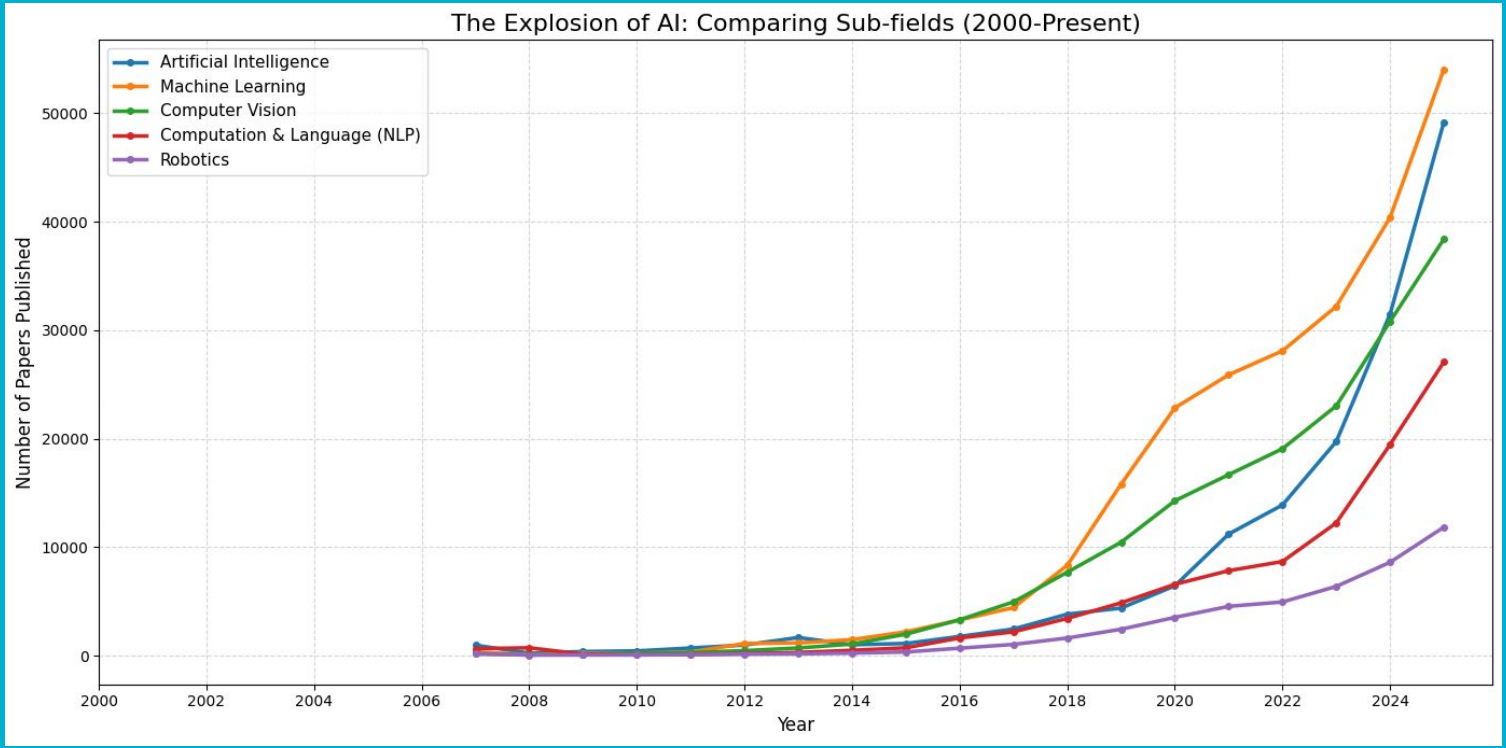
Misspelling Search Comparison

 # Misspelled Query: retrival augmnted generatoin models

- TF-IDF Search : 5/100
- BM25 and Vector Search : 20/100
- Neural Network Reranking : 25/100

Conclusion: With heavy typos, TF-IDF collapses, the hybrid model partially recovers semantic meaning, and the neural reranker picks the best among imperfect candidates, but true RAG papers never appear because upstream retrieval failed.

Historical Growth of Major AI Subfields

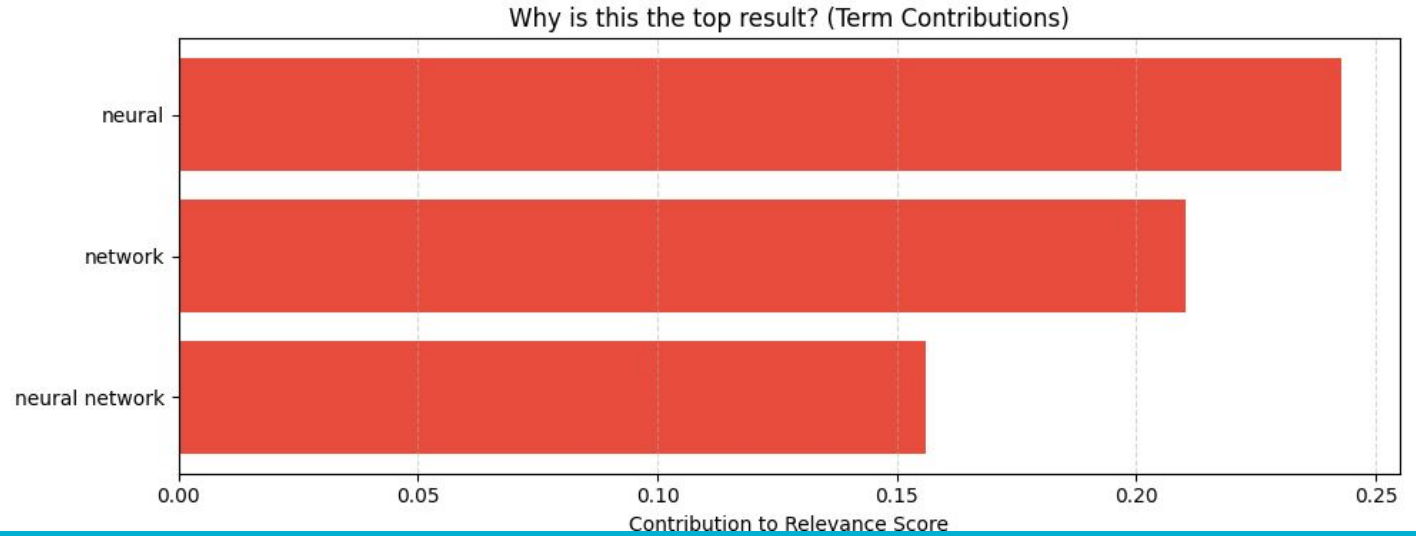


Insight: highlights the relative dominance of specific fields like **Computer Vision** and **Machine Learning**

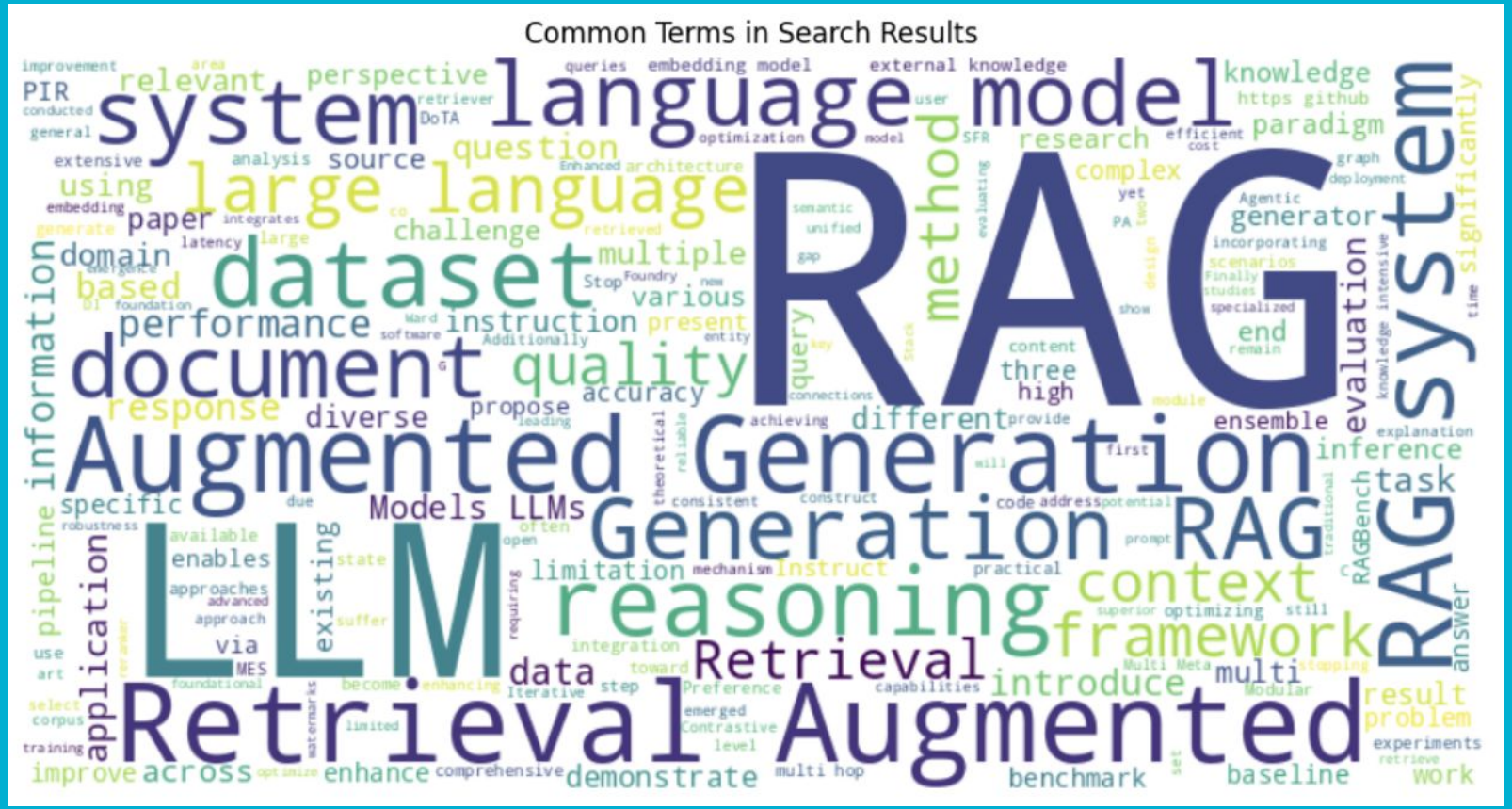
Query: neural network Max Results: 8 Search

```
Searching for 'neural network'...
```

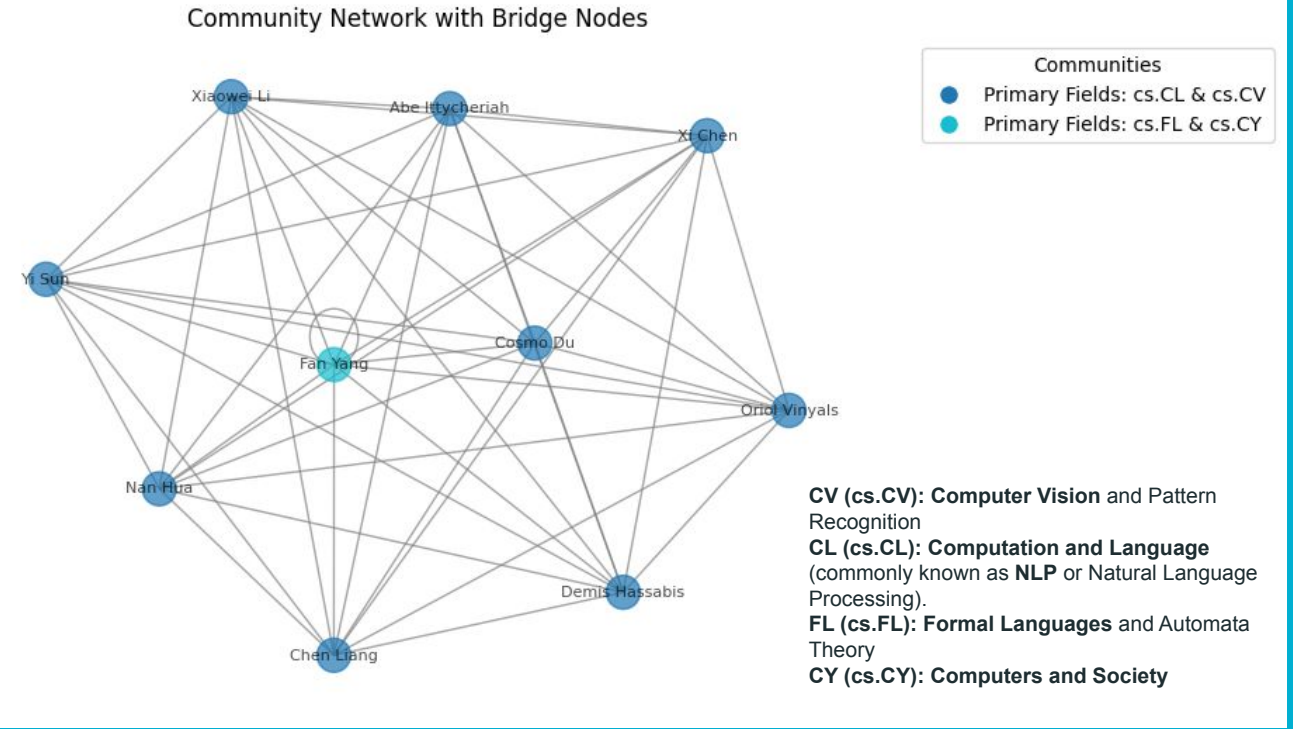
	title	authors	update_date	relevance_score
23072	Assessing Intelligence in Artificial Neural Ne...	Nicholas J. Schaub, Nathan Hotaling	2020-06-05	0.609067
10717	A Study on Neural Network Language Modeling	Dengliang Shi	2017-08-25	0.602459
60924	Guaranteed Quantization Error Computation for ...	Wesley Cooke, Zihao Mo, Weiming Xiang	2023-04-28	0.599723
12183	Visualizing Neural Network Developing Perturba...	Yadong Wu, Pengfei Zhang, Huitao Shen and Hui ...	2018-08-08	0.597041
103898	Interpreting Neural Networks through Mahalanob...	Alan Oursland	2024-10-28	0.579572
149967	The tip-of-the-tongue phenomenon: Irrelevant n...	Petro M. Gopych	2007-05-23	0.577609
102923	PiLocNet: Physics-informed neural network on 3...	Mingda Lu, Zitian Ao, Chao Wang, Sudhakar Pras...	2025-06-24	0.569150
56093	Safety Verification of Neural Network Control ...	Weiming Xiang and Zhongzhu Shao	2023-01-19	0.556966



Highlighting the most Dominant terms across the Papers



Social Structure of the Research Field by visualizing Author Collaborations



CV (cs.CV): Computer Vision and Pattern Recognition
CL (cs.CL): Computation and Language (commonly known as **NLP** or Natural Language Processing).
FL (cs.FL): Formal Languages and Automata Theory
CY (cs.CY): Computers and Society

Interpretation: Nodes bridging **cs.CV** and **cs.CL**, such as **Cosmo Du** and **Chen Liang** drive intersectional innovation like visual question answering

-> Interdisciplinary Hubs: Highlights key authors bridging isolated fields (e.g., AI & Robotics)
-> Structural Clarity: Uses colour-coded clusters to instantly reveal landscape fragmentation

Score (logits) Distribution of Cross-Encoder