THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Scholarly Topic Navigator

Explainable Research Digest Pipeline

Team LexiCore: Aditya Kanbargi, Pramod Krishnachari, Trisha Singh

December 2025

The Information Overload Crisis

The Problem

Thousands of NLP/ML papers flood platforms like ArXiv and ACL monthly. Researchers struggle to keep up, often spending 20+ hours weekly just tracking new publications.

Manual Discovery

Time-consuming and incomplete, leading to missed critical research.

Black Box Systems

Current recommendation engines lack transparency on *why* a paper is suggested.

Our Solution: Explainable Pipeline

We introduce an automated pipeline that not only aggregates and processes research but explains its recommendations.

Aggregate & Preprocess

Collects papers from ArXiv, ACL, and S2ORC, applying rigorous NLP cleaning.

Embed & Retrieve

Uses Word2Vec, SBERT, and SciBERT for deep semantic understanding and hybrid search.

Explainability

The key differentiator: LIME-based interpretable classification tells users *why* papers matter.

End-to-End System Architecture

A modular pipeline designed for scalability and transparency.

Data Ingestion

APIs fetch data from ArXiv, ACL Bibtex, and S2ORC, followed by normalization.

Preprocessing

Language detection, tokenization, lemmatization, and stopword removal.

Embeddings

Vectorization using Word2Vec (100d), SBERT (384d), and SciBERT (768d).

Applications

Hybrid retrieval, Zero-Shot classification, and LIME explainability.

Dashboard

Streamlit interface for search, analytics, and visual explanations.

Primary Data Sources

ArXiv (arxiv.org)

Preprint server providing 20,983 papers across CS, AI, and ML categories. 100% abstract availability.

ACL Anthology

Official conference repository. Parsed 118,461 entries, though limited by lack of abstracts in Bibtex format.

Semantic Scholar

Academic search engine. Collected 1,553 papers with abstracts via Graph API, despite rate-limiting challenges.

Data Collection Statistics

From over 140k raw entries to a clean, usable dataset.

141K

Raw Collected

Total entries from all three sources.

22.5K

Final Clean Dataset

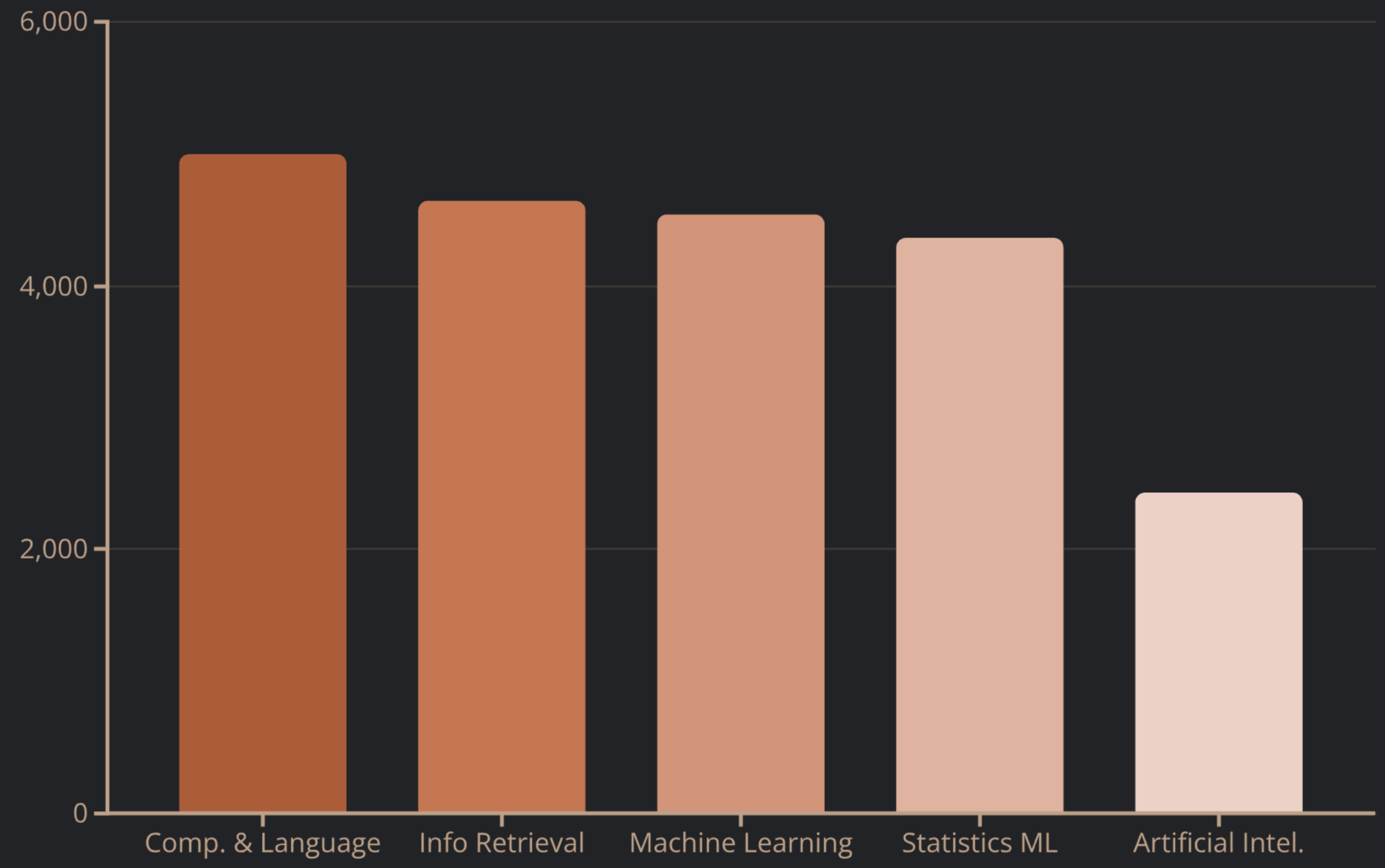
Papers with valid abstracts retained.

4K

Duplicates Removed

Cross-category overlap eliminated.

ArXiv Category Distribution



Data Validation & Cleaning

Ensuring high-quality input for the NLP models required a rigorous filtration process.

1

Validation Checks

Identified 82.9% missing abstracts (mostly ACL) and 2,832 missing authors. Validated year range (1995-2025).

2

Filtering Criteria

Retained only English papers with titles >10 chars and abstracts >50 chars. Must have at least one author.

3

Deduplication Strategy

Prioritized peer-reviewed sources: ACL › S2ORC › ArXiv. Kept the first occurrence based on this hierarchy.

Text Preprocessing Pipeline

Transforming raw text into structured data took approximately 14 minutes for 22,522 papers.

1. ### Language Detection

   Filtered out non-English papers (Indonesian, French, Catalan) using `langdetect`.

2. ### Text Cleaning

   Removed URLs, emails, and normalized whitespace.

3. ### Special Char Removal

   Stripped LaTeX, unicode, and brackets while keeping basic punctuation.

4. ### Tokenization

   Segmented sentences (NLTK) and tokenized words (spaCy), removing stopwords.

5. ### Lemmatization

   Converted words to base forms and filtered tokens under 2 characters.

Preprocessing Statistics

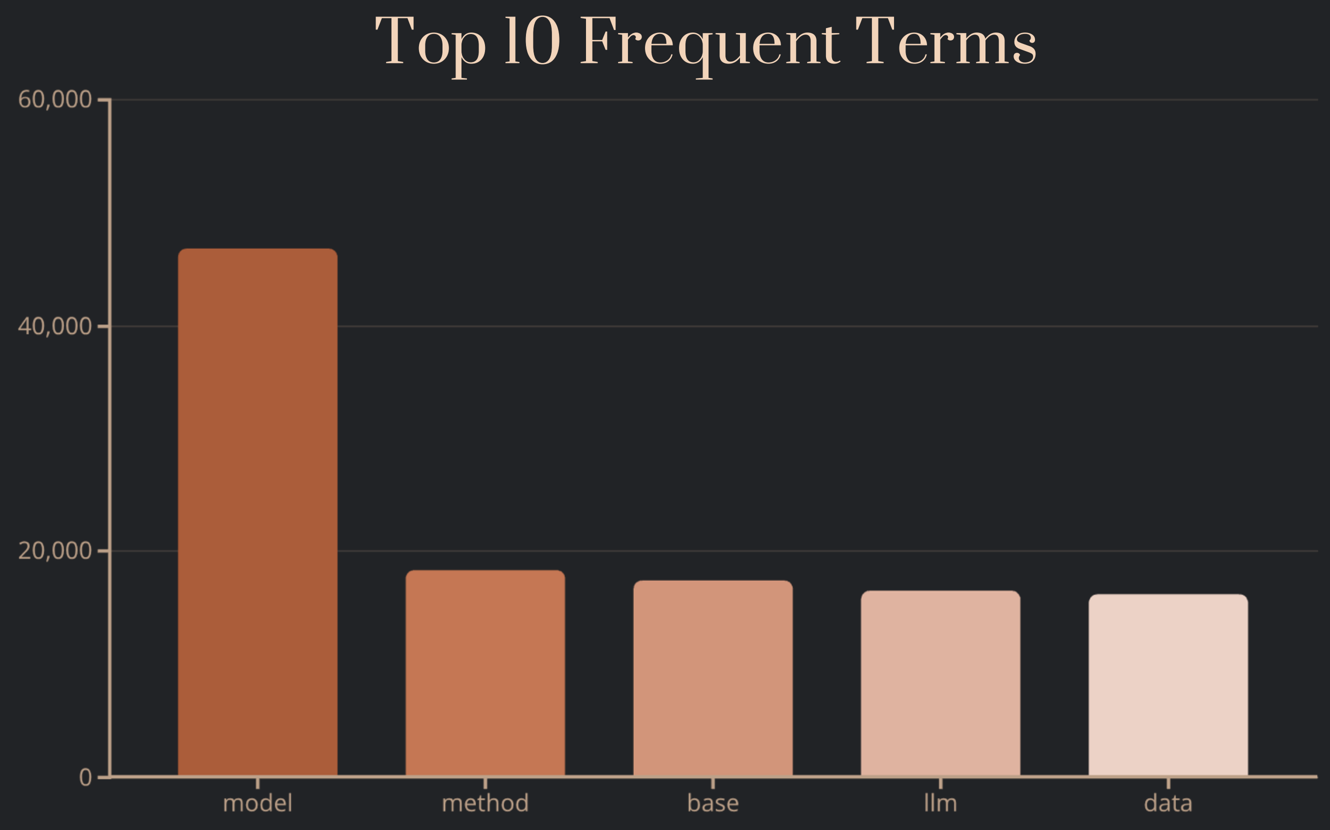
Token Metrics

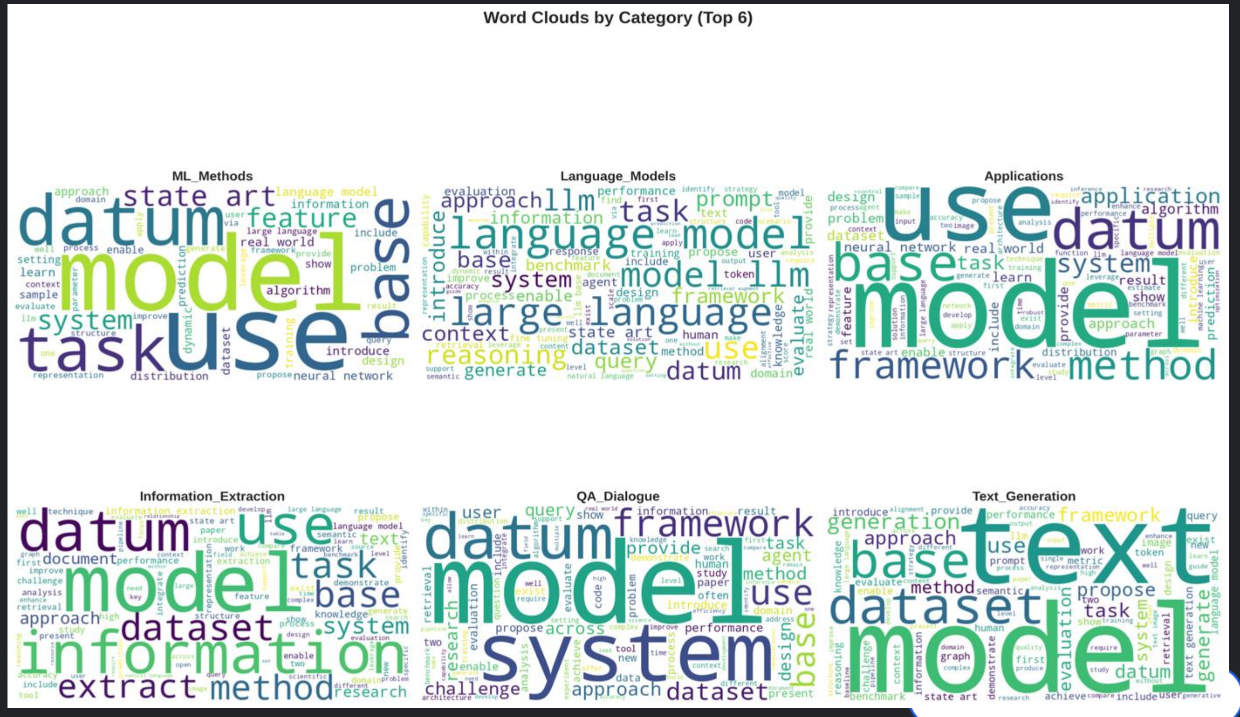
2.7M
Total Tokens

48K
Unique Tokens

121
Avg Tokens/Paper

The high frequency of "LLM" (Rank 4) reflects the dominant research trend of 2024-2025.





Embeddings, Retrieval & Classification

A technical deep dive into modern embedding representations, hybrid retrieval engines, and zero-shot classification pipelines for scientific literature analysis.

Embedding Model Architectures

We evaluated three distinct approaches to representing text as vectors, ranging from traditional baselines to domain -specific transformers.

Word2Vec (Baseline)

A traditional Gensim implementation using Skip-gram/CBOW architecture.

- Dimensions: 100
- Vocab: 28,727 words
- Method: Average of word vectors

SBERT (General Purpose)

Modern `all-MiniLM-L6-v2` model trained on 1B+ sentence pairs.

- Dimensions: 384
- Speed: 5x faster than BERT-base
- Output: Abstract & Title embeddings

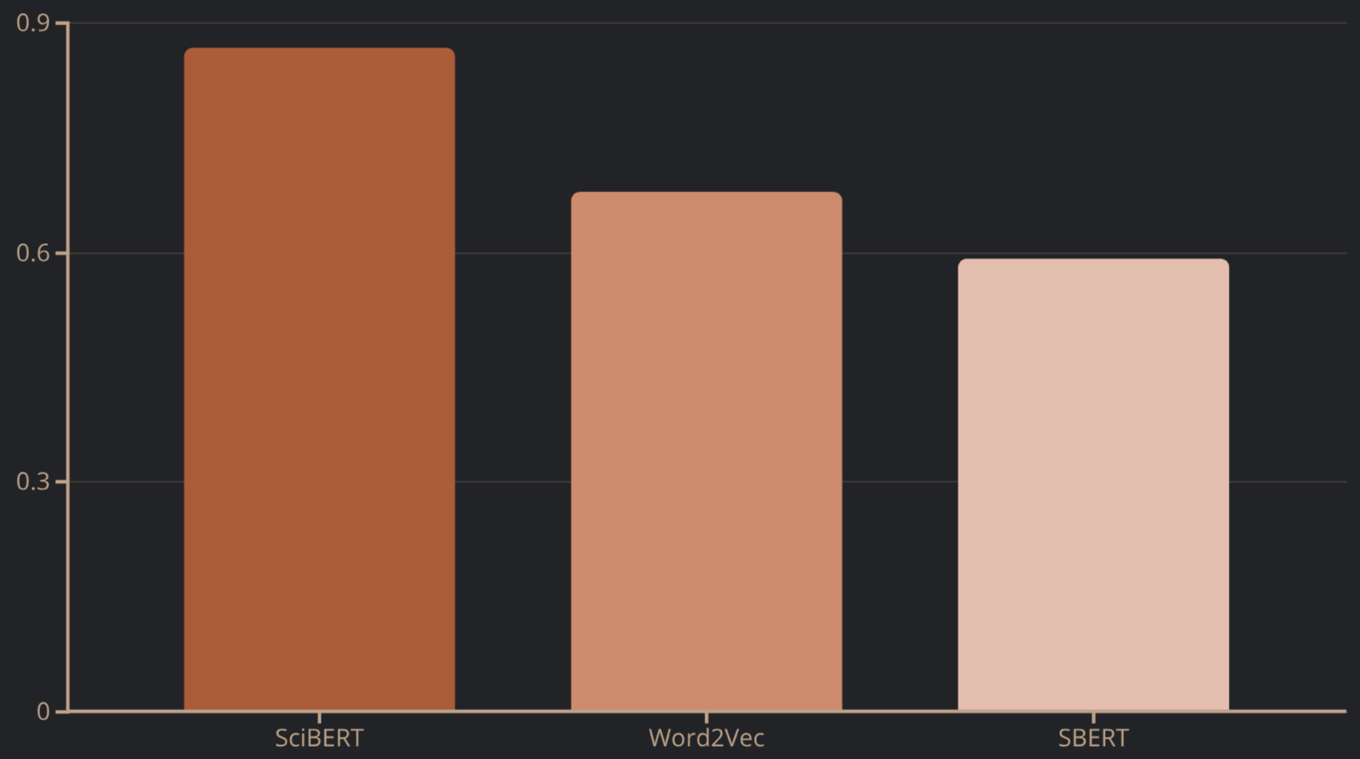
SciBERT (Domain Specific)

The `allenai-specter` model optimized for citation prediction.

- Dimensions: 768
- Params: 110 million
- Training: Scientific papers

Embedding Quality Analysis

We analyzed the Average Pairwise Cosine Similarity across a sample of 10 papers. Lower similarity indicates better separation for clustering tasks.



Assessment Results

- ### SBERT (Winner)

  Achieved the best separation (0.593), making it ideal for topic modeling.

- ### SciBERT Issues

  High similarity (0.867) indicates "semantic collapse," where all papers appear too similar, risking clustering failure.

- ### Recommendation

  Use SBERT for topic modeling; reserve SciBERT for specific scientific similarity searches.

Retrieval Engines: Keyword vs. Semantic

BM25 (Keyword Retrieval)

Traditional probabilistic retrieval based on term frequency and inverse document frequency.

- Algorithm: Okapi BM25 with saturation.
- Pipeline: Tokenization, Stopword removal, Porter stemming.
- Strength: Exact keyword matching.

FAISS (Vector Search)

Facebook AI Similarity Search for efficient dense vector retrieval.

- Index: IndexFlatIP (Exact search via Inner Product).
- Process: L2 Normalization of SBERT embeddings.
- Strength: Captures semantic meaning beyond keywords.

Key Difference: BM25 finds exact text matches (e.g., "transformer"), while FAISS finds conceptual matches (e.g., "attention mechanism" results for "transformer" queries).

Hybrid Retrieval System

To achieve the best of both worlds, we implemented a fusion algorithm using Weighted Reciprocal Rank.

1. Dual Retrieval

Execute parallel queries: get exact matches via BM25 and semantic matches via FAISS.

2. Normalization

Normalize scores from both engines to a [0, 1] scale to ensure comparability.

3. Weighted Fusion

Combine scores: $\alpha \times BM25 + (1-\alpha) \times Semantic$.

4. Re-Ranking

Final results are re-ranked based on the combined score.

Configuration: We utilize a semantic weight of 0.7 and a keyword weight of 0.3. This prioritizes semantic understanding to handle synonyms and varied scientific terminology.

Zero-Shot Classification

We utilized the `facebook/bart-large-mnli` model (400M parameters) to label 5,000 papers without training data. The model classifies content into 12 defined NLP research areas.

ML Methods

Language Models

Applications

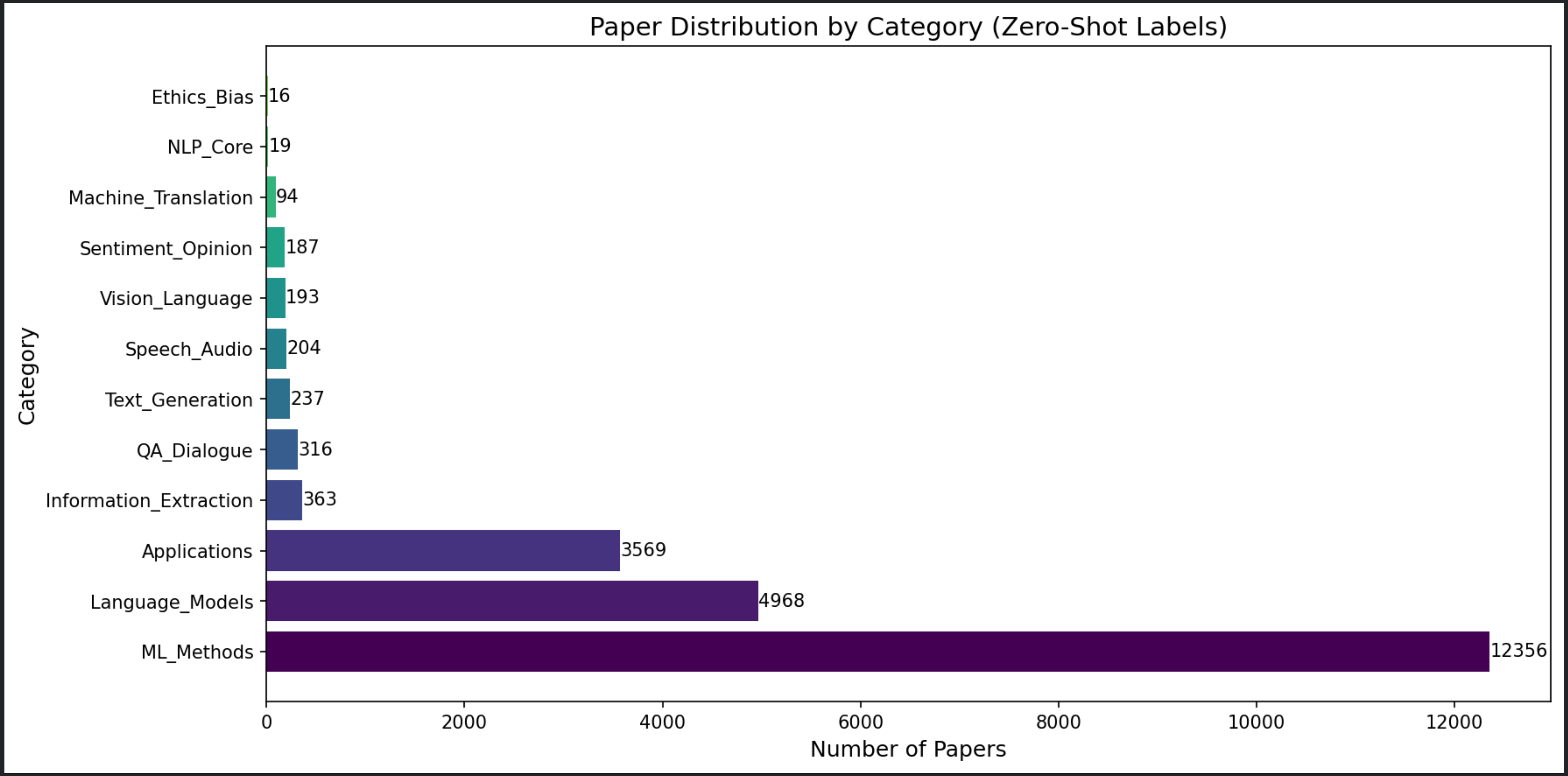
Info Extraction

QA & Dialogue

Text Generation

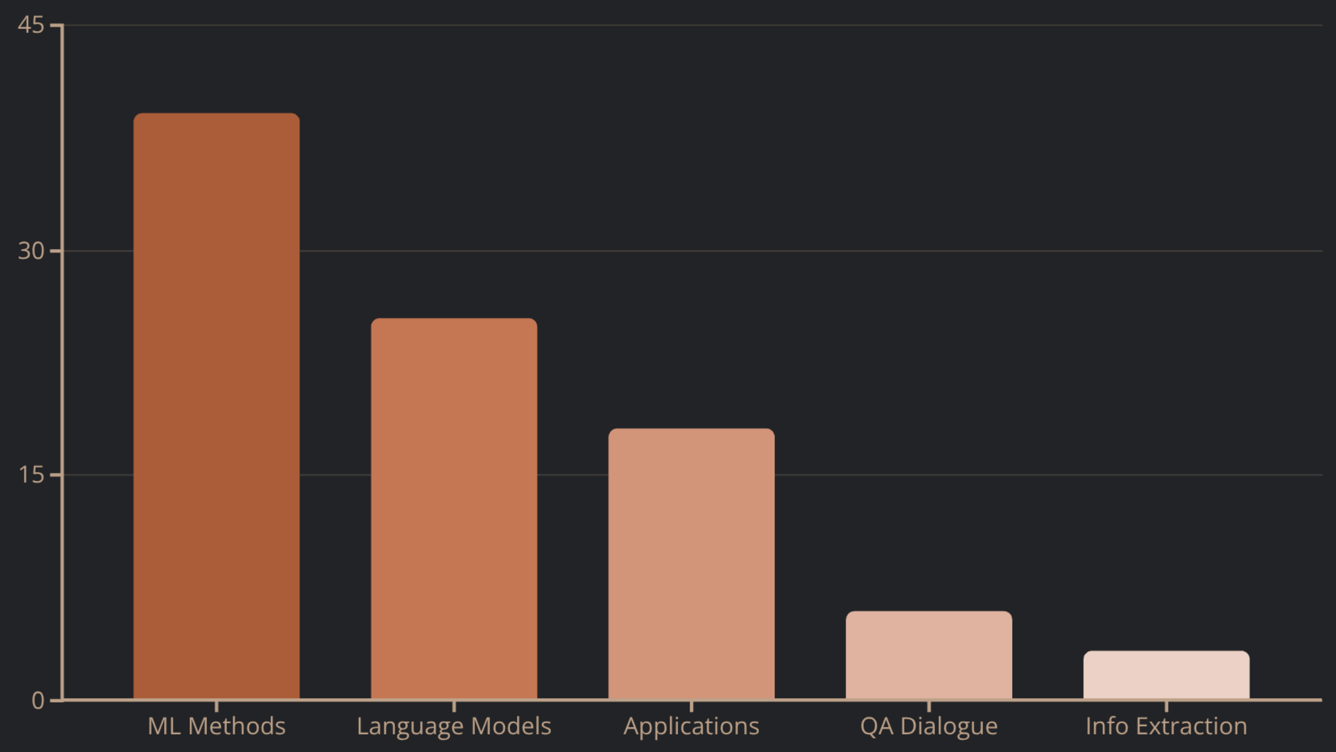
Speech/Audio

Vision Language



Classification Distribution Results

Analysis of 5,000 sampled papers reveals the current landscape of NLP research.



Key Insights

- ML Methods Dominance (39%): Reflects the heavy volume of algorithmic research in ArXiv's cs.LG category.
- LLM Boom (26%): Language Models represent the second largest category, driven by the 2024-2025 research surge.
- Practical Focus (18%): Significant portion of research is dedicated to real-world applications.

Production Pipeline: Classification & Summarization

Supervised Classifier

For production speed, we trained a lightweight classifier on the zero-shot pseudo-labels.

01

Input

Paper abstract text.

02

Encoder

Facebook's BART-large-MNLI model

03

Model

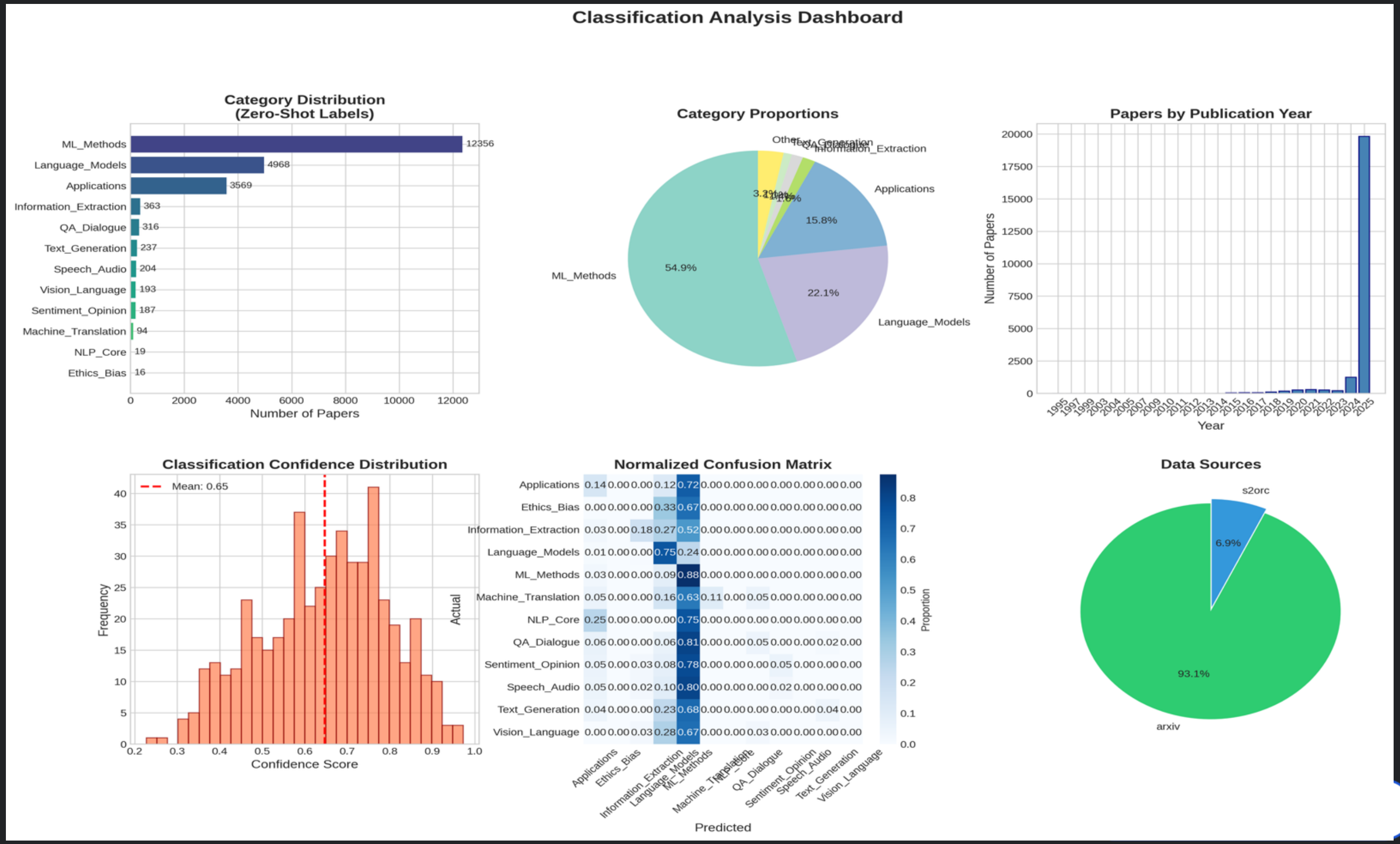
Logistic Regression.

Advantage: Fast inference and interpretable weights.

Summarization Engine

We employ a dual approach depending on the user need.

- Extractive (TextRank): Logistic regression algorithm that selects top sentences. Very fast; used for quick previews.
- Abstractive (Facebook's BART-large-MNLI model): 400M parameter model that generates novel paraphrased text. Slower; used for detailed digests.



Unsupervised Topic Discovery

Beyond zero-shot classification, we employed unsupervised topic modeling to uncover latent thematic patterns within the research corpus. This approach reveals natural topic clusters without predefined categories.

LDA (Latent Dirichlet Allocation)

Configuration: 10 topics using Variational Bayes

Approach: Assumes documents are mixtures of topics, where each topic represents a distribution over words. Iteratively refines assignments through probabilistic inference.

Tradeoffs: Fast and interpretable, but bag-of-words representation loses contextual information and word order.

BERTopic (Neural Topic Modeling)

Pipeline: SBERT embeddings → UMAP dimensionality reduction → HDBSCAN clustering → c-TF-IDF representation

Model: all-MiniLM-L6-v2 (384-dimensional semantic vectors)

Approach: Embeds documents in semantic space, reduces dimensions while preserving structure, applies density-based clustering, then extracts representative words using class-based TF-IDF.

Tradeoffs: Captures semantic relationships and synonyms effectively but requires more computational resources and careful hyperparameter tuning.

Topic Discovery: Results & Evaluation

The BERTopic processing pipeline leverages advanced techniques to extract meaningful topics from unstructured text data.

Text Corpus

Raw documents are preprocessed.

SBERT Encoding

Semantic embeddings are generated.

UMAP Reduction

High-dimensional data is compressed.

HDBSCAN

Density-based clustering identifies topics.

Topic Words

Representative keywords are extracted.

Coherence Evaluation

Evaluating the quality of discovered topics using coherence scores:

- LDA Coherence Score: 0.42
- BERTopic Coherence Score: 0.58

Key insight: Contextual embeddings in BERTopic capture semantic nuances more effectively, leading to higher coherence scores compared to traditional methods like LDA.

Discovered Topics: Sample Results

0	machine, learning, models, data	Core ML concepts & techniques
1	neural, networks, deep, computer, vision	Advanced deep learning architectures
2	natural, language, processing, text, generation	NLP and text-based applications
3	reinforcement, agent, environment, reward	Reinforcement learning theory & practice
4	robotics, control, motion, autonomous	Robotics and autonomous systems
5	big, data, cloud, distributed, computing	Scalable data processing & cloud
6	ethics, bias, fairness, responsible, AI	Ethical considerations in AI development
7	healthcare, medical, diagnosis, imaging	AI applications in medicine & health
8	finance, trading, risk, market	AI in financial modeling & markets
9	education, learning, student, personalized	AI's role in educational technology

Explainable NLP: Paper Discovery & Analysis

An end-to-end pipeline for scholarly paper discovery, featuring hybrid retrieval, zero-shot classification, and LIME-based explainability.

Explainability with LIME

We utilize Local Interpretable Model-agnostic Explanations (LIME) to bring transparency to black-box models, ensuring users can trust recommendations in academic settings.

01

Perturb Data

Generate samples by removing or masking words from the input text.

02

Predict

Get model predictions for all perturbed variations.

03

Fit Linear Model

Fit an interpretable linear model locally to the data.

04

Extract Weights

Identify specific words that contributed positively or negatively to the prediction.

LIME in Action: Analyzing Predictions

How the model interprets specific text inputs for classification.

Input Text

> "Recent coreference resolution models rely heavily on span representations to find coreference links between word spans..."

Prediction: ML_Methods (Confidence: 89.72%)

Word Importance Analysis

Supports Prediction

coreference (+0.0001)

Opposes Prediction

links (-0.0001)

Streamlit Dashboard Features

A comprehensive interactive web application designed for researchers.

Paper Search

Natural language queries using BM25, Semantic, or Hybrid retrieval methods.

Analytics

Dataset statistics, category distributions, and embedding comparison metrics.

Explainability

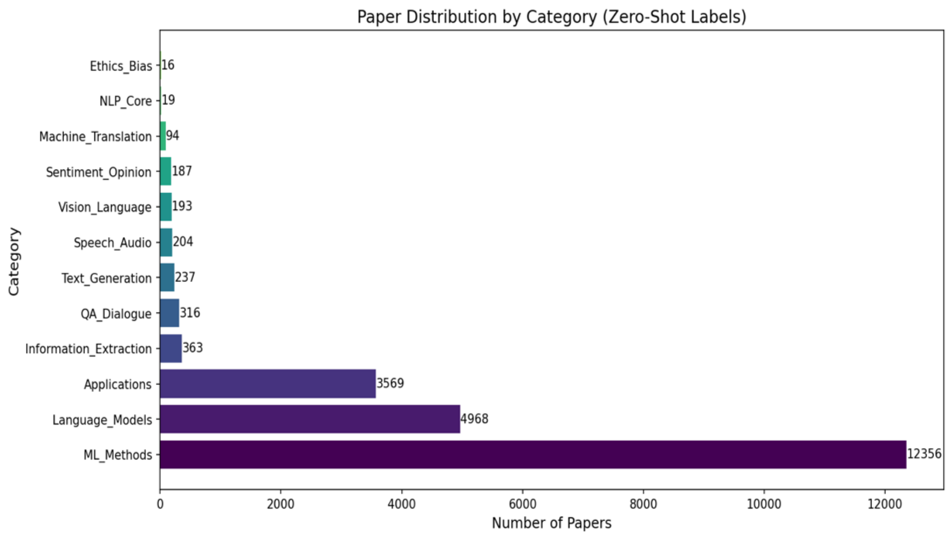
LIME explanations for predictions with word importance visualization.

Summarization

Dual approach using Extractive (TextRank) and Abstractive (BART) methods.

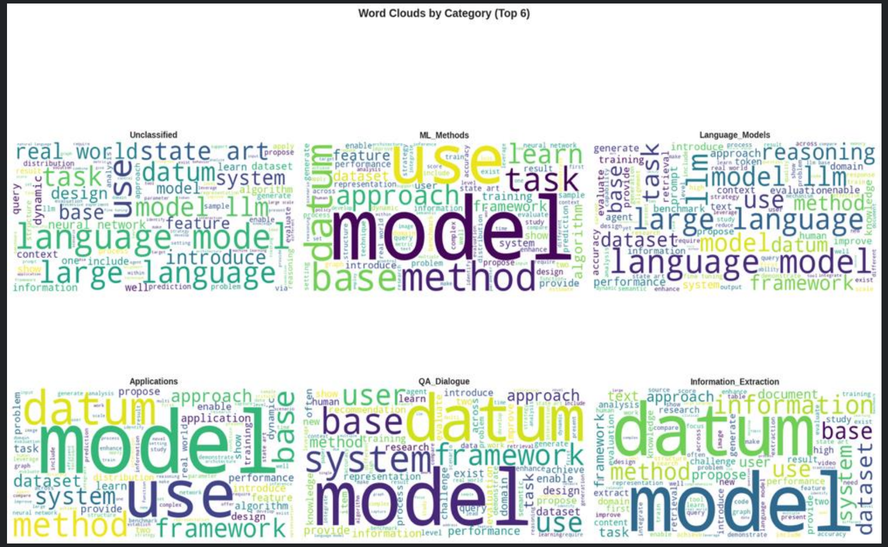
Visualization Gallery

Visual insights generated to analyze the dataset and model performance.



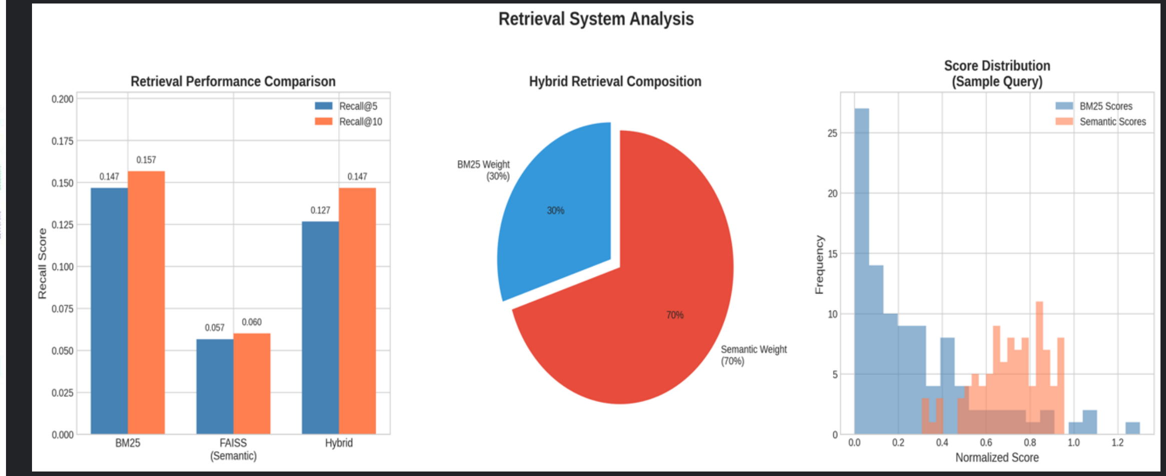
Category Distribution

Horizontal bar charts color-coded by frequency showing papers per category.



Word Clouds

2x3 grid visualizing top terms for each classification category.



Retrieval Dashboard

Analysis of BM25 vs FAISS performance, score distributions, and overlap.

Evaluation Metrics

Rigorous testing across the entire data and modeling pipeline.

1

Data Quality

Achieved 99.98% English language accuracy and removed 3.0% duplicates.

2

Embeddings

Evaluated embedding quality via mean intra class cosine similarity. SBERT achieved reasonable similarity score of 0.593 across papers within same category

3

Retrieval

Measured via Recall@k, Precision@k, and Mean Reciprocal Rank (MRR).

4

Classification

Zero-shot baseline accuracy assessed via per-category precision, recall, and F1 scores.

Key Results & Achievements

Summary of the data pipeline, retrieval system, and user interface performance.

22k	100ms	12
Papers Processed	Query Latency	Categories
Aggregated from 3 sources with 48k vocabulary tokens extracted.	Fast hybrid search combining BM25 and Semantic retrieval.	Defined and labeled using zero-shot classification on 5,000 papers.

Challenges & Solutions

Overcoming technical hurdles during development.

Missing Abstracts

Problem: ACL Anthology Bibtex lacked abstracts.Solution: Filtered during cleaning; future scraping planned.

API Rate Limiting

Problem: Semantic Scholar returned 429 errors.Solution: Implemented exponential backoff with jitter.

SciBERT Collapse

Problem: High similarity (0.867) caused poor clustering.Solution: Switched to SBERT for topic modeling tasks.

Library Conflicts

Problem: Numpy 2.0 broke NLP libraries.Solution: Pinned specific versions (numpy<2.0.0).

Conclusions & Future Work

A comprehensive academic search system combining explainability, hybrid retrieval, and zero-shot categorization to transform research discovery.

22K	48K	3	12
Research Papers	Unique Tokens	Embedding Models	Research Categories
Processed from three authoritative sources	Complete vocabulary coverage	W2V, SBERT, SciBERT compared	Automated via zero-shot classification

Explainability-First Design

LIME integration reveals exactly why papers are recommended, building trust and transparency in academic research settings.

Hybrid Retrieval Architecture

Combines BM25 keyword precision (30%) with FAISS semantic understanding (70%) for optimal search results.

Zero-Shot Taxonomy

Eliminates manual labeling requirements, enabling scalable classification across new research domains automatically.

Future Roadmap

Short-Term Enhancements

- ACL abstract scraping expansion
- Fine-tuning with human-labeled data
- User feedback integration loop
- Enhanced SciBERT clustering accuracy

Long-Term Vision

- Citation network graph analysis
- Personalized recommendation engine
- Real-time daily paper ingestion
- Multi-language support
- Zotero and Mendeley integration

Thank You