Research Lens

Automated Research Trend Analysis & Discovery Platform

Team Members

Rajas Panikar 25M2126	Shivraj Naghera 25M2136	Iqbal 25M2112
Built the frontend and connected it with backend API endpoints through REST APIs. Created Backend API and handled data operations with an ORM and created visual analytics dashboards.	Implemented algorithms to identify top-N trending keywords across scraped papers and perform intelligent analysis on the aggregated data. Database management and system design.	Developed a scraper to collect research papers and set up a Celery-based cron system to run background scraping periodically.

Problem Statement

Researchers and students are overwhelmed by the sheer volume of new papers published daily on platforms like arXiv.

Manually filtering through thousands of abstracts to find relevant, high-impact, or trending topics is inefficient and prone to missing critical developments.

There is a lack of centralized tools that combine automated scraping, intelligent summarization, and trend visualization in a single, offline-capable interface.

Scope

Management

Local PDF downloading, organization, and search functionality.

User Interface

A responsive web dashboard for filtering, searching, and viewing insights.

Automated Data Collection

Scheduled scraping of research repositories (primarily arXiv) for specific domains (CS, AI, Physics, etc.).

Intelligent Analysis:

Extraction of metadata, keyword frequency analysis, and NLP-based summarization of abstracts.

Visualization

Interactive graphs showing keyword velocity (trends over time) and topic distribution.

Limitations

1. **Source Restriction**

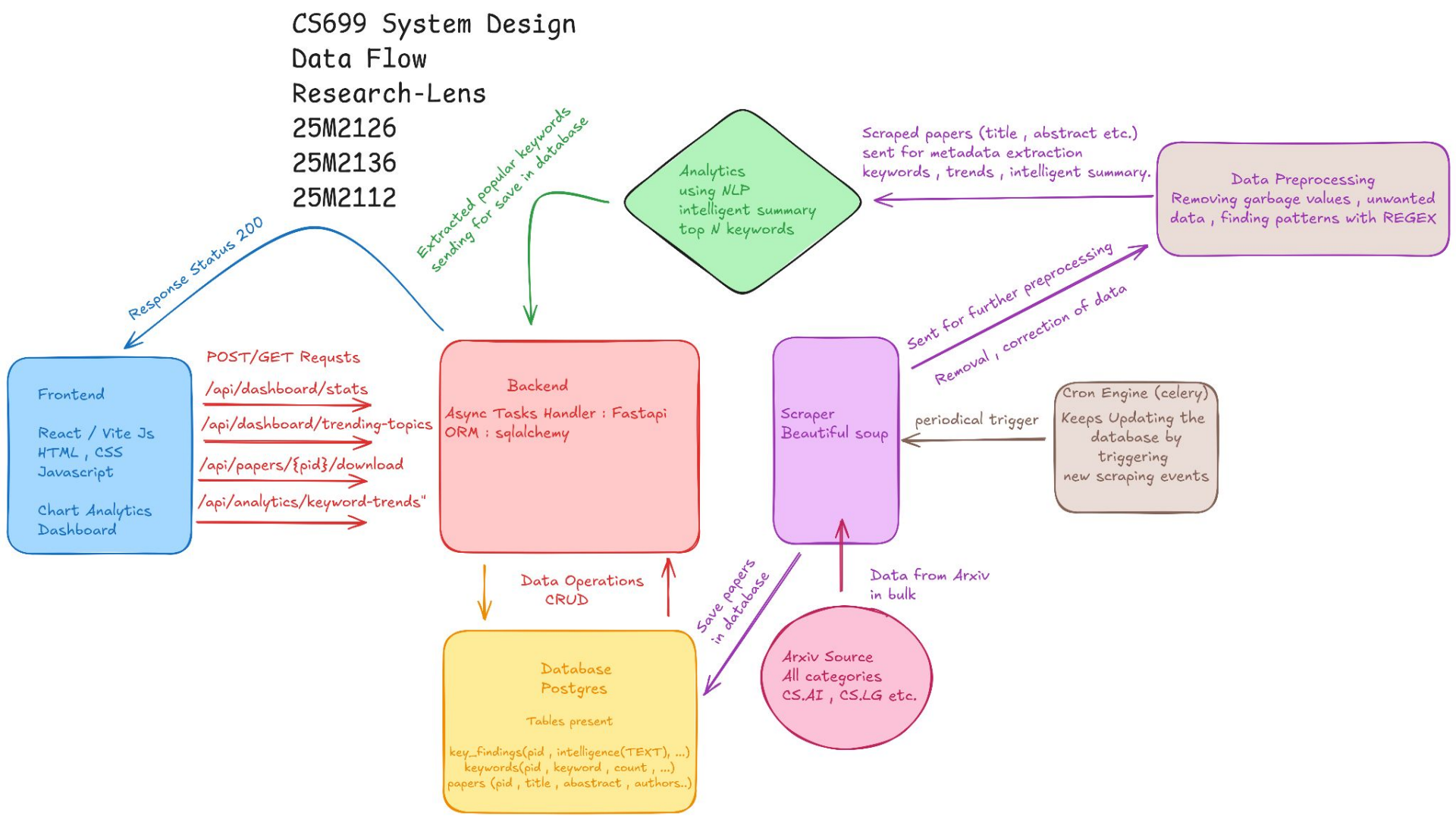
Currently optimized for arXiv; scraping other repositories (IEEE, Springer) is limited by paywalls and strict anti-bot measures.

2. **NLP Depth**

Summarization is based on abstracts, not full-text analysis, to maintain performance.

3. **Rate Limiting**

Data collection speed is throttled to comply with repository usage policies (e.g., arXiv's API limits).


CS699 System Design
Data Flow
Research-Lens
25M2126
25M2136
25M2112
Extracted popular keywords sending for save in database
Analytics using NLP intelligent summary top N keywords
Scraped papers (title , abstract etc.) sent for metadata extraction keywords , trends , intelligent summary.
Data Preprocessing
Removing garbage values , unwanted data , finding patterns with REGEX
Response Status 200
Frontend
React / Vite Js
HTML , CSS
Javascript
Chart Analytics Dashboard
POST/GET Requsts
/api/dashboard/stats
/api/dashboard/trending-topics
/api/papers/{pid}/download
/api/analytics/keyword-trends"
Backend
Async Tasks Handler : Fastapi
ORM : sqlalchemy
Sent for further preprocessing
Removal , correction of data
Scraper
Beautiful soup
periodical trigger
Cron Engine (celery)
Keeps Updating the database by triggering new scraping events
Data Operations
CRUD
Save papers in database
Data from Arxiv in bulk
Arxiv Source
All categories
CS.AI , CS.LG etc.
Database
Postgres
Tables present
key_findings(pid , intelligence(TEXT), ...)
keywords(pid , keyword , count , ...)
papers (pid , title , abastract , authors..)

Tech Stack

Frontend	Backend	Database	Scraping & Parsing	Analysis
React.js **Vite** **Recharts , matplotlib** (Visualization) **HTML** **CSS , Tailwind** (Custom Styling) **Lucide React** (Icons)	**Python** **FastAPI** **Uvicorn** (ASGI Server)	**PostgreSQL** (Data Storage) **SQLAlchemy** (ORM)	**Requests** **BeautifulSoup4** (HTML Parsing, **Regex** (Pattern Matching)	**Spacy** (Keyword Extraction) **Hugging Face Transformers** (en_core_web_md)

Contribution Details Frontend


```
rajas@rajas-Lenovo-ideapad-330S-14IKB:~/Desktop/CS699/labProject/Research-Lens-Frontend-2/research-
y=format:"%C(yellow)%h%Creset - %s %C(bold blue)<%an>%Creset" -n 30
bcfc9f4 - added trends page <rajaspaunikar>
c436e53 - added repository page <rajaspaunikar>
654aa6e - added dashboard page <rajaspaunikar>
d0c28c1 - basic react vite setp <rajaspaunikar>
rajas@rajas-Lenovo-ideapad-330S-14IKB:~/Desktop/CS699/labProject/Research-Lens-Frontend-2/research-
```

Contribution Details Backend


```
rajas@rajas-Lenovo-ideapad-330S-14IKB:~/Desktop/CS699/labProject/Research-Lens-Backend-Latest$ git log --pretty=format:"%C(yellow)%h%Creset - %s
 %C(bold blue)<%an>%Creset" -n 30
a53018b - Add files via upload <Panigala>
7d8c549 - first iteration <Panigala>
22b650a - auto scraping 2AM <Panigala>
caa3bbc - added env file for credential <Shivraj1906>
9d98c05 - added trend analyzer for trend analysis <Shivraj1906>
9fb49e4 - added metadata extractor for keyword extraction <Shivraj1906>
a3b6219 - mock data remove code fix <Shivraj1906>
3561b2a - Merge branch 'main' of github.com:rajaspaunikar/Research-Lens-Backend-Latest <rajaspaunikar>
8d891ec - task scheduler auto matic <Panigala>
9cc183c - removing mock data <Shivraj Naghera>
229e25c - scraper <Panigala>
33ddd07 - scraper tester <Panigala>
4f8a13f - scraper <Panigala>
72188cd - added main file endpoints <rajaspaunikar>
83ca7b3 - Create README.md <Panigala>
rajas@rajas-Lenovo-ideapad-330S-14IKB:~/Desktop/CS699/labProject/Research-Lens-Backend-Latest$
```

Project Details & Screenshots

Dashboard Details

Dashboard Stats

Displays
* Papers Parsed
* High Impact findings
* Recent Influx stats.

Visual

Representation

Includes
* Trending Keywords
* Leaderboard
* Latest Findings

PAPERS PARSED
992
TOTAL DATABASE

HIGH IMPACT
853
SOTA & NOVEL FINDINGS

RECENT INFLUX
+992
LAST 7 DAYS

TRENDING KEYWORDS

1	Llm	98 hits
2	Large Language Models Llm	72 hits
3	Large Language Model	63 hits
4	The-Art	39 hits
5	Image	37 hits

LATEST FINDINGS

SOTA 11/21/2025

In contrast, state-of-the-art vision approaches are increasingly building on large, pre-trained self-supervised models rather than training from scratch.

info 11/21/2025

In this work, we strive to improve LMM reasoning capabilities in a purely unsupervised fashion without any annotated data or reward distillation.

PERFORMANCE 11/21/2025

Experiments on challenging THuman2.0, XHuman, and HuGe100K data show that NoPo-Avatar outperforms existing baselines in practical settings without ground-truth poses and delivers comparable results in lab settings with ground-truth poses.

Repository Details

Papers Table	Visuals	Filter / Searching
A paginated table of all scraped papers. Includes search functionality (Title/Author) and a one-click "Download PDF" button that saves files locally.	Clean, tabular layout with status indicators for downloaded files. consectetur adipiscing elit.	Ability to search any word in the complete database on any field.

RESEARCH REPOSITORY (992 found)

Search...

ID	Title	Authors	Date	Actions
2511.16674v1	Dataset Distillation for Pre-Trained Self-Supervised Vision Models cs.CV cs.AI	George Cazenavette, Antonio Torralba...	11/21/2025	
2511.16672v1	EvoLMM: Self-Evolving Large Multimodal Models with Continuous Rewards cs.CV	Omkat Thawakar, Shravan Venkatraman...	11/21/2025	
2511.16673v1	NoPo-Avatar: Generalizable and Animatable Avatars from Sparse Inputs without Human Poses cs.CV	Jing Wen, Alexander G. Schwing...	11/21/2025	
2511.16671v1	Thinking-while-Generating: Interleaving Textual Reasoning throughout Visual Generation cs.CV cs.AI	Ziyu Guo, Renrui Zhang...	11/21/2025	
2511.16670v1	Learning to Think Fast and Slow for Visual Language Models cs.CV	Chenyu Lin, Cheng Chi...	11/21/2025	

Trend Analysis

A multi-line chart allowing users to compare the growth of specific keywords (e.g., "LLM" vs. "Diffusion") over the last 30 days.

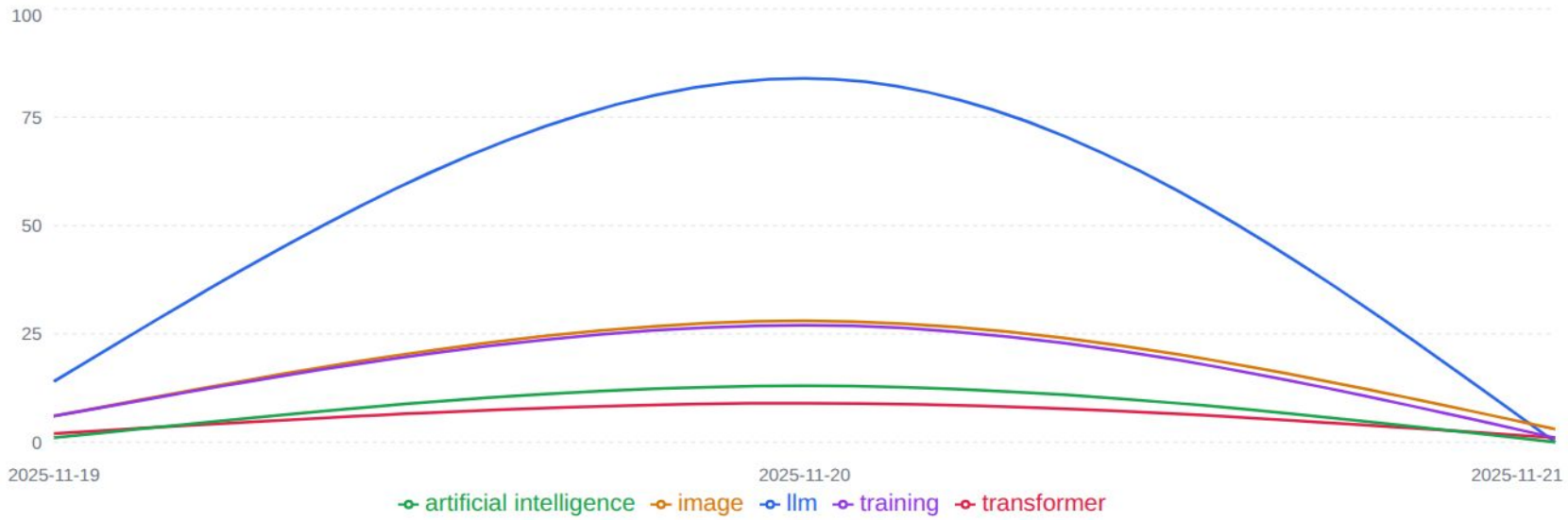
Keyword Velocity Analysis

e.g. diffusion

Add

Compare research interest over the last 30 days.

LLM × TRANSFORMER × ARTIFICIAL INTELLIGENCE × IMAGE × TRAINING ×



Challenges Faced

Making too many requests in a short time caused arXiv to block our scraper.

Solution: We implemented a "polite" scraping mechanism with time.sleep() delays between requests and rotated User-Agent headers to mimic legitimate browser traffic.

Loading 500+ papers into the React table caused the browser to lag.

Solution: We implemented Pagination (calculating pages on the client side) and Debouncing on the search input. This ensures the UI remains responsive even with large datasets.

arXiv search results display dates in inconsistent formats (e.g., "originally announced Dec 2023" vs. "submitted 12 May").

Solution: We wrote a robust Regex parser in the scraper that looks for specific patterns and falls back to a default date only if parsing fails entirely.

Future Enhancements

LLM Integration

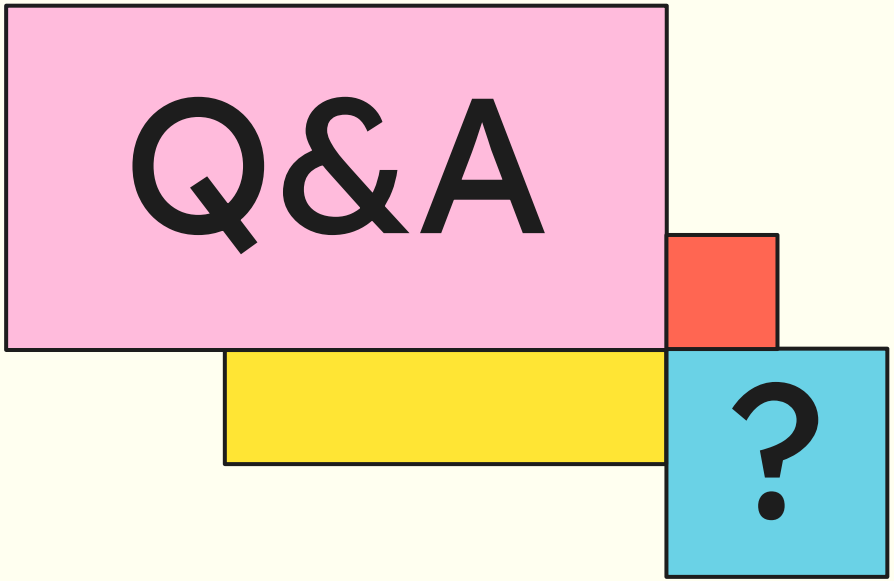
We plan to integrate a local LLM (like Llama 3) to generate "One-Line Takeaways" for every paper automatically.

Cross-Source Scraping

Extending the scraper to support Hugging Face Papers and Semantic Scholar.

User Accounts

Adding authentication so different users can save their own "Watchlists" of keywords.


Q&A
?