Part One: Interest Statements

Problem & Why We Care (NLP × Academic Search)

Researchers and students navigating vast digital academic libraries like arXiv often struggle to efficiently find the most relevant research. While traditional keyword-based searches work for simple queries, they fail to capture contextual intent or related concepts in more complex queries, making it difficult for users to discover truly relevant or related papers.

We want to build a state-of-the-art semantic search system specifically for arXiv, leveraging advanced language model embeddings and efficient vector search. Our goal: enable fast, relevant discovery, reduce researcher frustration, and foster easier exploration of new ideas. As a team, we're committed to improving research workflows and democratizing knowledge access through modern NLP.

Team Interests:

- Son: Interested in optimizing models for highly efficient, domain-tuned text embeddings.
- Kyle: Focused on implementing robust and scalable search pipelines.
- Daniel: Keen on evaluating system effectiveness via user studies targeted at researchers.
- Jinbo: Energized by building high-performance search indexing and frontend design.

Part Two: Landscape Analysis

Item	Description	Commentary
Semantic Scholar	AI-powered academic search with text embeddings.	Strong semantic search baseline for papers; industry reference point.
Google Scholar	Mainstream keyword-based academic search.	Vast user base, but lacks deep semantic search capabilities.
FAISS	Efficient open-source vector search library.	Industry standard for high-performance semantic indexing.
Arxiv Sanity Preserver	Community tool for fast arXiv search with text similarity.	Demonstrates user demand for advanced academic search workflows
SciBERT	Transformer model pretrained on scientific papers.	Foundational for domain-specific (especially ML/CS) embedding and search.
DenseViz	Visualization tool for dense vector (embedding) spaces.	Useful for debugging and developing embedding-based search systems.

OpenSearch with Vector Plugin	Open source, scalable search engine with encoding and retrieval for vector search.	Practical production backend for text embeddings, enterprise-ready and flexible.
PaperDigest	AI-powered paper summarization.	Potential for later integration, but not the focus for core semantic search.
Awesome Semantic Search	Curated repo of semantic search tools, papers, projects (github.com/Agrover112/awesome-semantic-search).	One-stop resource hub for system design, benchmarks, and code examples.
Searchthearxiv	Open-source semantic search engine for arXiv ML papers using OpenAI embeddings and Pinecone.	Highlights practical embedding-based search but reveals room for deeper, context-aware exploration.

Part Three: Problem Statement

Researchers face inefficiency and fatigue when discovering and understanding relevant scientific work on arXiv, due to keyword search limitations and the challenge of surfacing related or contextually appropriate papers. This slows research progress and knowledge sharing—especially in rapidly evolving scientific domains.

Part Four: Proposed Solution

We propose "ArXplorer," an academic search assistant for arXiv that:

- Utilizes state-of-the-art transformer-based semantic text embeddings (e.g., SciBERT or similar).
- Employs efficient vector search (FAISS, OpenSearch) to return contextually relevant results based on meaning, not just keywords.
- Integrates citation and keyword metadata to boost ranking and exploratory features.
- Features a clean, researcher-focused interface supporting query refinement, filters, and interactive exploration.

Part Five: High-Level Technical Approach

- Text Embeddings: Use or fine-tune transformer-based models trained on scientific texts for strong semantic representations.
- Query Understanding: Transformer embeddings for user intent understanding, typo robustness, and synonymy.
- Vector Search Backend: Implement vector similarity search with FAISS or OpenSearch for rapid matching of queries to papers.
- UI/UX: Build an intuitive interface for input, result exploration, and filters (date, author, keyword, citation).
- Evaluation: Run pilot studies with target users to measure relevance, satisfaction, and impact.

Part Six: Milestones (8-week Timeline)

- W1:

    - Set up repo & development environment; acquire and preprocess arXiv data (title, abstracts, metadata).
    - Benchmark and select base embedding model (e.g., SciBERT/SBERT/Specter/Specter2).
- W2:
    - Implement embedding pipeline; generate text embeddings for papers.
    - Build and test initial FAISS/OpenSearch vector index.
    - Develop baseline semantic search interface.
- W3:
    - Integrate search query embedding and similarity-based retrieval.
    - Build minimal frontend for submitting and displaying search queries and results.
- W4 (Demo Milestone):
    - Demo initial search system—allow natural language queries and return top-ranked papers.
    - Gather internal feedback for further development and quality tweaks.
- W5:
    - Integrate citation/keyword metadata for ranking and filters.
    - Improve UI (filters, autocomplete, sorting options).
- W6:
    - Optimize performance, ensure scalability.
    - Prepare for user-facing pilot testing.
- W7:
    - User study for evaluation of search quality, usability, and relevance on target audience.
- W8:
    - Finalize features and documentation.
    - Prepare project report, demo video, and plan next-phase enhancements (e.g., expanded indexing, advanced ranking).

Part Seven: Unknowns & Risks to Investigate

- Domain adaptation: Ensuring embeddings remain tuned for the fast-evolving language of scientific publication.
- Large corpus scalability: Efficient retrieval with millions of papers.
- Evaluation: Establishing rigorous search relevance metrics for academic use cases.

- UI design: Presenting semantic search results in a way that fosters exploration and learning.

Date: September 26, 2025

HEADLINE

We announce ArXplorer, a semantic search engine designed to revolutionize how researchers discover and understand academic papers on arXiv.

SUB-HEADING

ArXplorer combines cutting-edge transformer-based embeddings and fast vector search for effortless, context-aware scientific exploration.

SUMMARY

ArXplorer indexes and semantically understands millions of arXiv papers, enabling researchers to perform refined, intent-driven searches. The system aims to reduce cognitive overload, accelerate discovery, and support multidisciplinary knowledge synthesis.

PROBLEM

Navigating vast academic repositories like arXiv is hindered by imprecise keyword searches and difficulty extracting contextually relevant papers—slowing research progress.

SOLUTION

ArXplorer leverages advanced language model embeddings and high-performance vector search to provide accurate, contextually relevant, and easily explorable search results.

QUOTE FROM LEADERSHIP

“ArXplorer reflects our passion for using the latest advances in language technology to empower researchers with tools for faster, more insightful discovery,” said Son Nguyen, project lead.

HOW TO GET STARTED

Users can try ArXplorer via a simple web interface or desktop app. Queries can be submitted using natural language for immediate, relevant scientific paper recommendations.

CUSTOMER QUOTE

“As a PhD student, ArXplorer helped me uncover relevant papers that traditional search engines continually overlooked,” said Kyle, early user.

CLOSING AND CALL TO ACTION

ArXplorer will be available soon. Visit www.arxplorer.ai to sign up for early access and experience next-generation academic search.

Appendix: Focused Messaging

Student Edition

- Headline: “ArXplorer: Your research companion for smarter paper discovery.”
- Subhead: “Semantic search for academic success.”
- Summary: Empowers students to discover relevant research quickly and confidently.

Professional/Lab Edition

- Headline: “ArXplorer: Redefining academic literature search for professionals.”
- Subhead: “Precision, speed, and insight through AI-powered semantic search.”
- Summary: Targets research labs seeking productivity and depth in literature reviews.

Institutional/Library Edition

- Headline: “ArXplorer empowers academic digital libraries.”
- Subhead: “Enhance your collection’s value with robust, context-aware search.”
- Summary: Appeals to libraries and institutions upgrading research infrastructure.

References to Tools and Further Reading:

- [SciBERT model & guide]
- [Awesome Semantic Search curated list]
- [Sentence Transformers (SBERT)]
- [FAISS documentation]
- [OpenSearch vector search]
- [Searchthearxiv open-source engine]
- [Guide: How to Build a Semantic Search Engine in Python]
- [Systematic paper on semantic search in scientific repositories]
- [Intelligent Semantic Search for Academic Journals]